BIOGENESIS Jurnal Ilmiah Biologi

pISSN 2302-1616, eISSN 2580-2909
Vol 8, No. 1, June 2020, pp. 41-48
Available online http://journal.uin-alauddin.ac.id/index.php/biogenesis
DOI https://doi.org/10.24252/bio.v8i1.12002

Performance Comparison of Data Sampling Techniques to Handle Imbalanced Class on Prediction of Compound-Protein Interaction

AKHMAD REZKI PURNAJAYA[1*], WISNU ANANTA KUSUMA[2], MEDRIA KUSUMA DEWI HARDHIENATA[3]

[1]Department of Software Engineering, Faculty of Computer, Universal University
Sungai Panas, Batam, Indonesia. 29444
*Email: rezki.purnajaya@uvers.ac.id
[2]Tropical Biopharmaca Research Center, Faculty of Math and Science, IPB University
Jl. Taman Kencana, RT.03/RW.03, Bogor, West Java, Indonesia. 16128
[3]Department of Computer Science, Faculty of Mathematics and Natural Science, IPB University
Jl. Meranti Wing 20 Level 5 Kampus IPB Darmaga, Bogor, Indonesia. 16680

Received 7 January 2020; Received in revised form 8 March 2020;
Accepted 2 May 2020; Available online 30 June 2020

ABSTRACT

The prediction of Compound-Protein Interactions (CPI) is an essential step in the drug-target analysis for developing new drugs as well as for drug repositioning. One challenging issue in this field is that commonly there are more numbers of non-interacting compound-protein pairs than interacting pairs. This problem causes bias, which may degrade the prediction of CPI. Besides, currently, there is not much research on CPI prediction that compares data sampling techniques to handle the class imbalance problem. To address this issue, we compare four data sampling techniques, namely Random Under-sampling (RUS), Combination of Over-Under-sampling (COUS), Synthetic Minority Over-sampling Technique (SMOTE), and Tomek Link (T-Link). The benchmark CPI data: Nuclear Receptor and G-Protein Coupled Receptor (GPCR) are used to test these techniques. Area Under Curve (AUC) applied to evaluate the CPI prediction performance of each technique. Results show that the AUC values for RUS, COUS, SMOTE, and T-Link are 0.75, 0.77, 0.85 and 0.79 respectively on Nuclear Receptor data and 0.70, 0.85, 0.91 and 0.72 respectively on GPCR data. These results indicate that SMOTE has the highest AUC values. Furthermore, we found that the SMOTE technique is more capable of handling class imbalance problems on CPI prediction compared to the remaining three other techniques.

Keywords: area under curve; compound-protein interaction; drug-target analysis; imbalanced class; SMOTE

INTRODUCTION

The identification of Compound-Protein Interaction (CPI) plays a key role in the development of drugs, particularly herbal medicines. The great advances in molecular medicine and the human genome project provide more opportunities to discover unknown associations in the CPI network. The new interactions that are discovered can be helpful for finding new drugs by screening candidate compounds and also essential to understand the causes of side effects in existing drugs (Mei *et al*., 2013; Hong *et al*., 2017). Currently, the latest computational models have been discovered in predicting of potential compound-protein interactions, including deep learning techniques (Tsubaki *et al*., 2019).

However, at this moment, there are only a few studies available to understand the interaction between compounds and proteins. For example, PubChem and ChEMBL database store 90 million drug candidate compound records, but some compounds interaction to protein targets are still limited (Wang *et al*., 2017; Mendez *et al*., 2019). The computational method for predicting the CPI is thus essential in drug or herbal medicine studies. The method can reduce time, cost, and failure rate for discovering new drugs or herbal medicines (Kim *et al*., 2013).

To address the above issue, some studies on CPI predictions have been conducted by Biopharmaca Research Centre in Bogor, Indonesia. Indonesia Jamu Herbs (IJAH) webserver is developed by Biopharmaca Research Center to predict the efficacy of herbal of drug formulas for various diseases using the multicomponent-multitarget network

that consists of plant-compound interaction, compound-protein interaction, and protein-disease association networks (Masri & Kusuma, 2018). There are many medicinal properties of herbal formula, which cannot be predicted by IJAH due to a lack of CPI data. To solve this problem, a previous study by Kurnia (2017) has predicted CPI in IJAH by using the Bipartite Local Model–Neighbor Interaction profile Inferring (BLMNII). BLMNII has a good ability to predict new compounds or new protein data, which has a non-interacting pair (Kurnia, 2017). Also, BLMNII can solve the problem of other pharmacological network prediction that predicts LncRNA-Disease Associations (Cui *et al*., 2019) and Biomedical Bipartite Networks (Zhang *et al*., 2020). However, the study by Kurnia (2017) has not solved the class imbalance problem in the prediction of CPI. Another problem that may occur when an algorithm is created while ignoring data balance is that the prediction might be biased towards the majority class while ignoring the minority class (Chawla, 2003).

To overcome the imbalanced class in CPI, a study to compare CPI prediction performance by using Random Under-sampling (RUS) and Balanced Sampling techniques (Mousavian *et al*., 2016). Mousavian *et al*. gave some results from experiments in 2016 that the RUS technique has better results than Balanced Sampling. Ezzat *et al.* has also conducted another relevant study in 2016 by evaluating CPI prediction using Synthetic Minority Oversampling Technique (SMOTE). This is done by incorporating the Decision Tree. The Decision Tree has initially shown lower performance in predicting CPI than the Support Vector Machine (SVM). The study has also demonstrated that SMOTE implemented with a Decision Tree had better prediction performance than only using SVM. Then, an experiment has proven Tomek-Link (T-Link) can improve performance in the classification of arterial blood pressures and Ecoli2 datasets (Elhassan *et al.*, 2017). Based on those three studies, we conclude that RUS, SMOTE, and T-Link techniques are proper sampling techniques to handle the imbalanced class on CPI.

Besides using the sampling techniques mentioned above, we try to implement a Combination of Over-Under-sampling (COUS) technique to handle the class imbalance problems in CPI prediction. COUS is done by balancing the amount of data distribution by increasing the amount of minor class data (oversampling) and reducing major class data (undersampling). However, after the matrix of CPI has been balanced by using the data sampling technique, the CPI matrix might have missing values of interacting class caused by duplication or reduction. To overcome this problem, we use k-Nearest Neighbors (k-NN) to impute missing values. This approach can be easily adjusted to work with any attribute as a class, using only distance metrics to modify attributes. This approach can also efficiently treat examples with multiple missing values (Batista & Monard, 2002).

This study used two Yamanishi datasets (i.e., Nuclear Receptor and G-Protein Coupled Receptor), a common benchmark dataset on CPI prediction. We then compare four data sampling techniques, i.e., RUS, Combination of Over-Under-sampling (COUS), SMOTE, and Tomek Link (T-Link); see the effectiveness of the technique to handle class imbalance problem on CPI prediction. To handle missing values when conducting sampling data, we implemented k-Nearest Neighbour imputation. We use the Bipartite Local Model (BLM) as CPI prediction method was first introduced by Bleakley & Yamanishi (2009) and improved by combining BLM and Hubness-Aware Regression in Buza & Peška (2017). BLM create two local models using SVM as a classifier. The CPI prediction result using the BLM method will then be evaluated by using the Area Under Curve (AUC) and Receiver Operating Characteristic (ROC) (Sonego *et al*., 2008). AUC is a numerical measure to differentiate model performance and can be employed to show how successful the model rankings are by separating positive and negative observations. AUC is known to have proven to be a reliable performance measure for class imbalance problems (Fawcett, 2004).

MATERIALS AND METHODS

Datasets. This study used two of four Yamanishi datasets, Nuclear Receptor and G-Protein Coupled Receptor (GPCR), which are benchmark datasets on CPI prediction (Yamanishi *et al*., 2008). These datasets were downloaded from http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/. The nuclear Receptor dataset consists of 54 compounds, 26 proteins, and 1404 compound-protein interaction pairs that comprised 1314 non-interacting and 90 known interacting pairs. The GPCR dataset consists of 223 compounds, 95 proteins, and 21185 compound-protein interaction pairs that comprised 20550 non-interacting and 635 interacting pairs.

Data sampling techniques. To see the effectiveness of several techniques for solving this problem, we compare four data sampling techniques: RUS, COUS, SMOTE, and T-Link, which will be discussed in the following subsections.

In RUS, data from classes with a large number of instances (majority class) are removed randomly. The selection and removal processes were repeated until the majority class is equal to the minority class (Mousavian *et al.*, 2016). Firstly, the number of difference between the minority class and the majority class is calculated as follows:

$$Mean = \left(\frac{nMajority + nMinority}{2} \right).$$

Then, $\Delta majority$ as the number of differences between the majority class and the mean, $\Delta minority$ as the number of differences between the minority class and the mean is calculated as follows:

$$\Delta majority = nMajority - mean$$
$$\Delta minority = nMinority - mean.$$

Next, we remove the data of the majority class as many as randomly. After that, we duplicate the data of the minority class as many as randomly.

SMOTE works by creating synthetic data, i.e., replication data from minor data. SMOTE method works by searching k-NN for every single data in a minor class. After that, it makes synthetic data as much as the desired duplication percentage between minor data and k-NN, chosen randomly. SMOTE method is known to avoid overfitting when synthetizing minority class data (Chawla *et al*., 2002). Illustration of the SMOTE is shown in the Figure 1.

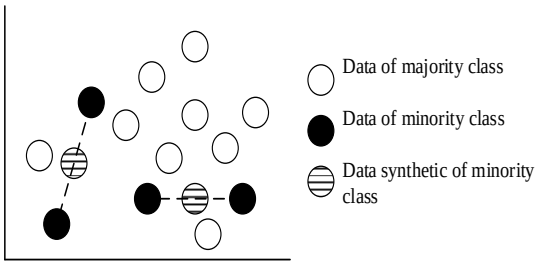

Figure 1. Illustration of the SMOTE technique (Hu & Li, 2013)

The T-Link algorithm was defined as a refinement of the Condensed Nearest Neighbor (CNN) technique, where CNN could choose the subset from all classes using One-Nearest Neighbor (1-NN). It only reduced data on the majority class that has been done 1-NN because if it reduces the minority class again, it will add the probability of misclassification later. For example, *x_i* and *x_j*, where the minority class (*x_i*) ≠ majority class (*x_j*) created a T-link pair and generated the *x_k* sample. The new *x_j* is reduced by *x_k* (Elhassan *et al*., 2017).

Missing data Imputation. CPI prediction runs if every compound and protein already has an interaction class. Therefore, to fill the values of NA on the CPI matrix, data imputation is needed. Data imputation is a technique that can be used to estimate the value of missing data by obtaining a pattern of data that has full features (Batista & Monard, 2002).

In this study, we use k-NN imputation to fill the missing interaction class. We can implement a k-NN imputation by following the following steps. First, the data was loaded and initialized the value of k for k-NN. For getting the predicted class, iterate from 1 to a total number of missing interaction class. Then the distance was calculated between the test data and each row of training data. Here, we use Gower distance as our distance metric. We then sort the calculated distances in ascending order based on distance values. Next, top k rows can be obtained from the sorted array and the most frequent class of these rows. Finally, missing interacting class is filled by predicted class.

Prediction. We use the Bipartite Local Model (BLM) algorithm and SVM classifier to predict CPI. The BLM algorithm was first proposed by Bleakley and Yamanishi (2009) and it has recently been shown to be effective in predicting CPI. The algorithm is as follows. First of all, the first local model denoted as $Model^1$ is first studied based on the interaction profile *Iji* and the protein similarity matrix SIM_{ii}^{p}. The equation to calculate *Model1* is as follows:

$$Model^1 = ClassifierTrain(SIM_{ii}^{p}, I_{ji})$$

where *Iij* is compound-protein interaction matrix, *i* is the index of the protein, and *j* is the index of the compound. Then, predict pre_{ij}^{1} by testing $Model^1$ with SIM_i^p as the *i-th* row of the protein similarity matrix. The prediction is calculated as follows:

$$pre_{ij}^{1} = ClassifierTest(Model^1, SIM_i^{p}).$$

The next step is to create the second local model as $Model^2$. This is done by training a classification algorithm based on the interaction profile *Iij* and the compound similarity matrix SIM_{jj}^{c} as follows:

$$Model^2 = ClassifierTrain(SIM_{jj}^{c}, I_{ij}).$$

Then, predict by testing with as the j-th row of the compound similarity matrix as follows:

$$pre_{ji}^{2} = ClassifierTest(Model^2, SIM_j^{c}).$$

Finally, the prediction results (pre_{ij}) are obtained by taking the maximum value of pre_{ji}^{2} and a transpose of pre_{ji}^{2} as follows:

$$pre_{ij} = \max(pre_{ij}^{1}, \left(pre_{ji}^{2}\right)^{T}).$$

We use 10-fold cross-validation to evaluate the performance of SVM on BLM. Cross-validation was one of the methods used to measure the stability of SVM for predicting testing data.

To measure CPI performance, the ROC curve is visualized, as shown in Figure 2. If the curve is more likely to go to the upper left corner, then it can be ascertained that the CPI prediction result can solve the class imbalance problem because it classifies precisely the positive class and the negative class data. Conversely, if the curve is closer to the baseline or the line across from (0, 0) point to (1, 1) point, then the data is not well classified because the data have an imbalance class (Sonego *et al.,* 2008).

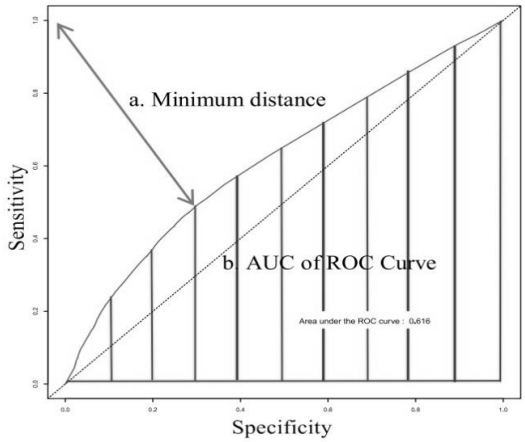

Figure 2. A basic ROC curve (Sonego *et al.,* 2008)

$$\text{Sensitivity} = \frac{TP}{TP + FN} \qquad \text{AUC} = \frac{Sensitivity + Specificity}{2}.$$

$$\text{Specificity} = \frac{TN}{TN + FP} \qquad \text{Accuracy} = \frac{TP+TN}{(TP+TN+FP+FN)}$$

Measured ROC parameters in this study were sensitivity, specificity, and AUC. From the equation above, the sensitivity and specificity values can be calculated from a confusion matrix. This table consists of TP (True-Positive), FP (False-Positive), FN (False-Negative), and TN (True-Negative) parts. After obtaining sensitivity and specificity values, we calculated AUC and Accuracy values. ROC is made by plotting sensitivity value on the y-axis and specificity value on the x-axis, as shown in Figure 2.

After the performance prediction is obtained, the ratio of positive data (interacting data class) can be calculated to see the percentage increase in the ratio of positive data between training data and prediction data.

$$\text{Ratio of Positive Data} = \frac{n_1}{n_s \times n_p}$$

where *n1* is the number of interacting class, *ns* is the number of compounds, and *np* is the number of protein (Harris, 1967).

RESULT AND DISCUSSION

Figures 3 and 4 show the CPI prediction evaluation results using ROC parameters previously implemented by BLM and data sampling techniques (RUS, COUS, SMOTE, and T-Link) on two Yamanishi datasets, i.e., Nuclear Receptor and GPCR. It can be seen in Figures 3 and 4 that each CPI prediction evaluation on two Yamanishi datasets gives different AUC values. On the Nuclear Receptor dataset using RUS, COUS, SMOTE, and T-Link sampling techniques, the AUC values are 0.77, 0.75, 0.85, and 0.79, respectively, as in Figure 3.

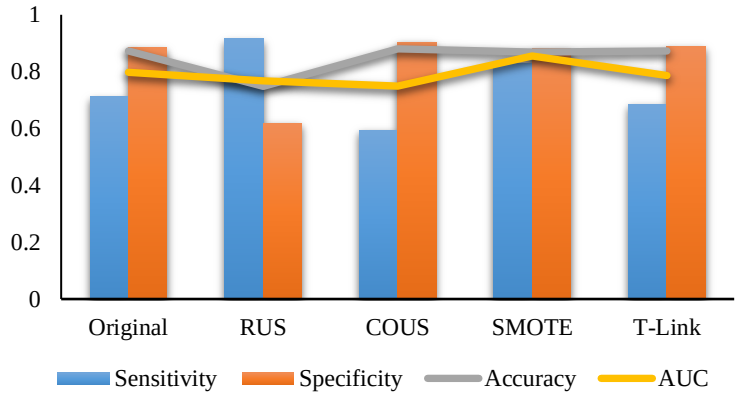

Figure 3. Statistics of CPI prediction performance on Nuclear Receptor dataset

On the other hand, the AUC values are 0.85, 0.70, 0.92, and 0.72, respectively, on the GPCR dataset, as shown in Figure 4. The higher the value of AUC, the more successful it is to distinguish model performance and separate positive and negative classes. The test result, which yields the largest AUC value of each Yamanishi dataset, is SMOTE. In this part, we found that the AUC values for SMOTE on Nuclear Receptor dataset is 0.85 and on the GPCR dataset is 0.92.

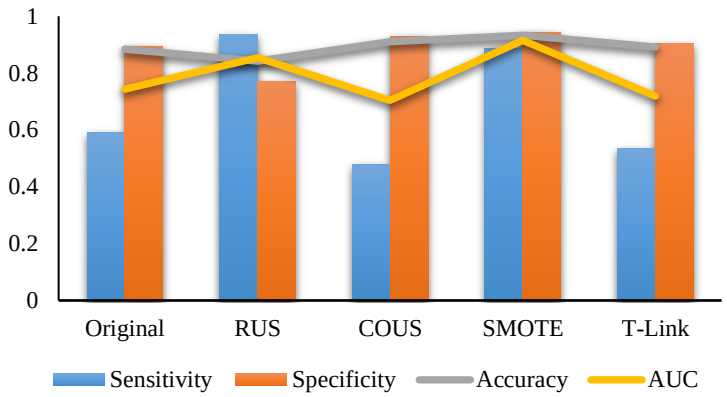

Figure 4. Statistics of CPI prediction performance on GPCR dataset

We also compare the AUC values from the original data with the AUC values in four data sampling techniques. As shown in Table 1, the AUC value on the Nuclear Receptor dataset for SMOTE is 0.05 higher than that of imbalanced data. Whereas on GPCR, RUS, and SMOTE datasets, the AUC values are 0.11 and 0.18 higher than imbalanced data.

Table 1. Value of CPI prediction performance

Dataset	Data Sampling Technique	Sensitivity	Specificity	Accuracy	AUC	The Difference of AUC with Imbalanced Data
Nuclear Receptor	Original	0.711	0.884	0.873	0.797	0
	RUS	0.915	0.619	0.748	0.767	0.030
	COUS	0.593	0.904	0.880	0.749	0.049
	SMOTE	0.830	0.880	0.868	0.855	0.057
	T-Link	0.686	0.888	0.873	0.787	0.011
GPCR	Original	0.592	0.894	0.885	0.743	0
	RUS	0.936	0.772	0.842	0.854	0.111
	COUS	0.480	0.928	0.910	0.704	0.039
	SMOTE	0.887	0.943	0.933	0.915	0.172
	T-Link	0.534	0.904	0.891	0.719	0.024

In addition, to compare AUC values of CPI prediction on each sampling technique, we display the ROC curve, which visualizes the performance of each data sampling technique for CPI prediction, as can be seen in Figure 5. In particular, Figure 5 shows the ROC curve of the predicted CPI on the Nuclear Receptor dataset in each sampling technique. The ROC curve of CPI prediction on the GPRC dataset with each sampling technique can be seen in Figure 6. It can be inferred from Figures 5 and 6 that the ROC curve of the SMOTE sampling technique is closer to (0.1) point than the ROC curves of other data sampling.

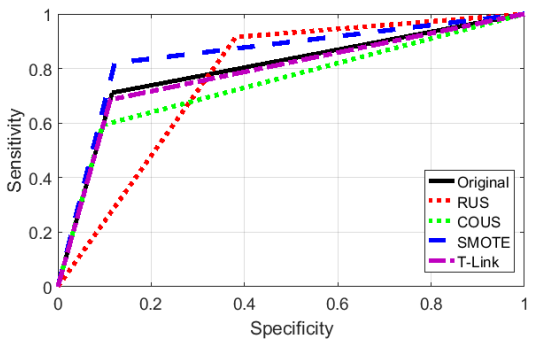

Figure 5. ROC curve of CPI prediction on Nuclear Receptor for each data sampling technique

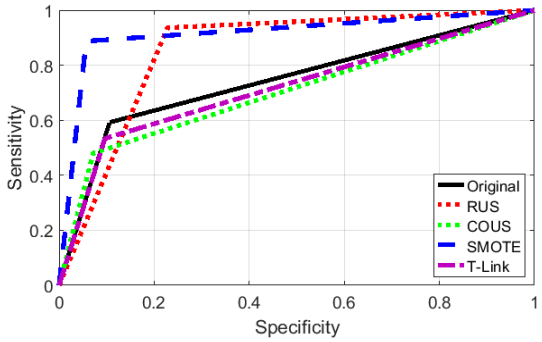

Figure 6. ROC curve of CPI prediction on GPCR for each data sampling technique

From AUC values and ROC curves, we conclude that the SMOTE data sampling technique can handle class imbalance problem in CPI predictions, better than the original data that contain imbalanced class or using the other sampling techniques such as RUS, COUS, and T-Link. RUS has better specificity values than other data sampling techniques, but the sensitivity values are much lower. RUS can predict interacting pairs better on other data sampling techniques. However, in this study, RUS cannot accurately predict the negative class. Hence RUS does not perform well enough to handle the imbalanced class in this case.

Table 2. Percentage increase in ratio of positive data

Dataset	Training Data			Prediction Data			Percentage increase in Ratio of Positive Data
	Interacting	Non-interacting	Ratio of Positive Data	Interacting	Non-interacting	Ratio of Positive Data	
Nuclear Receptor	90	1314	6.40%	317	1087	22.60%	16.20%
GPCR	635	20550	3%	4582	16603	21.60%	18.60%

The SMOTE can find new interacting pairs in CPI. This is evidenced by the increase in the percentage increase in the ratio of positive data by 16.2% in the Nuclear Receptor dataset and 18.6% in the GPCR dataset, as shown in Table 2.

CONCLUSION

We used four data sampling techniques: RUS, COUS, SMOTE, and T-Link, to balance the number of known interacting and non-interacting compound-protein pairs. In our experiments, SMOTE method had demonstrated better prediction performance than RUS, COUS, and T-Link techniques when 10-fold cross-validation was used. Also, we conclude that COUS and T-Link methods are unable to increase CPI prediction performance for an imbalanced class problem. Our experimental results also show that SMOTE has the highest AUC values, representing that it is reliable in sampling data and predicting interactions for new compounds or new protein data. In the future, there is a potential that SMOTE technique can be applied for CPI prediction, but it can also be used for drug-target interaction prediction, which also has a class imbalance problem. This technique can provide more information about new drugs and detect new targets for drug repositioning.

REFERENCES

Batista GEDAPA, Monard MC. 2002. A Study of K-Nearest Neighbour as an Imputation Method. *His.* vol 87: 251–260.

Bleakley K, Yamanishi Y. 2009. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics*. vol 25(18): 2397–2403. doi: https://doi.org/10.1093/bioinformatics/btp433.

Buza K, Peška L. 2017. Drug–target interaction prediction with Bipartite Local Models and hubness-aware regression. *Neurocomputing*. vol 260: 284–293. doi: https://doi.org/10.1016/j.neucom.2017.04.055.

Chawla NV. 2003. C4.5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure. In Proceedings of Workshop on Learning from Imbalanced Data Sets (II). August 21, 2003. Washington DC: ICML. vol 3: 66-73.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. vol 16: 321–357. doi: https://doi.org/10.1613/jair.953.

Cui Z, Liu JX, Gao YL, Zhu R, Yuan SS. 2019. LncRNA-disease associations prediction using bipartite local model with nearest profile-based association inferring. *IEEE Journal of Biomedical and Health Informatics*. vol 24(5): 1519–1527. doi: https://doi.org/10.1109/JBHI.2019.2937827.

Elhassan AT, Aljurf M, Al-Mohanna F, Shoukri M. 2017. Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rRUS) as a data reduction method. *Global Journal of Technology & Optimization*. vol S1: 1–11. doi: https://doi.org/10.4172/2229-8711.S1:111.

Ezzat A, Wu M, Li XL, Kwoh CK. 2016. Drug-target interaction prediction via class imbalance-aware ensemble learning. *BMC Bioinformatics*. vol 17(19): 267–276. doi: https://doi.org/10.1186/s12859-016-1377-y.

Fawcett T. 2004. ROC Graphs: Notes and practical considerations for data mining researchers. *Pattern Recognition Letters.* vol 31(8): 1–38.

Harris CW. 1967. Problems in measuring change. Madison: University of Wisconsin Press.

Hong M, Li S, Tan HY, Cheung F, Wang N, Huang J, Feng Y. 2017. A network-based pharmacology study of the herb-induced liver injury potential of traditional hepatoprotective Chinese herbal medicines. *Molecules.* vol 22(4): 1–14. doi: https://doi.org/10.3390/molecules22040632.

Hu F, Li H. 2013. A novel boundary oversampling algorithm based on neighborhood rough set model: NRSBoundary-SMOTE. *Mathematical Problems in Engineering*. vol 2013: 1–11. doi: http://dx.doi.org/10.1155/2013/694809.

Kim S, Jin D, Lee H. 2013. Predicting drug-target interactions using drug-drug interactions. *PloS One*. vol 8(11): 1–12. doi: https://doi.org/10.1371/journal.pone.0080129.

Kurnia A. 2017. Prediksi formula jamu berkhasiat menggunakan teknik link prediction dari jejaring bipartite senyawa aktif dan protein. [Thesis]. Bogor: IPB University.

Masri VR, Kusuma WA. 2018. Pengujian Usability pada Ijah-Webserver dengan Menggunakan Metode Cognitive Walkthrough. [Skripsi]. Bogor: IPB University.

Mei JP, Kwoh, CK, Yang P, Li XL, Zheng J. 2013. Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics.* vol 29(2): 238–245. doi: https://doi.org/10.1093/bioinformatics/bts670.

Mendez D, Gaulton A., Bento AP, Chambers J, De Veij M, Félix E, Magariños MP, Mosquera JF, Mutowo P, Nowotka M, Gordillo-Marañón M, Hunter F, Junco L, Mugumbate G, Rodriguez-Lopez M, Atkinson F, Bosc N, Radoux CJ, Segura-Cabrera A, Hersey A, Leach AR. 2019. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research.* vol 47(1): 930–940. doi: https://doi.org/10.1093/nar/gky1075.

Mousavian Z, Khakabimamaghani S, Kavousi K, Masoudi-Nejad A. 2016. Drug–target interaction prediction from PSSM based evolutionary information. *Journal of Pharmacological and Toxicological Methods*. vol 78: 42–51. doi: https://doi.org/10.1016/j.vascn.2015.11.002.

Sonego P, Kocsor A, Pongor S. 2008. ROC analysis: applications to the classification of biological sequences and 3D structures. *Briefings in Bioinformatics*. vol 9(3): 198–209. doi: https://doi.org/10.1093/bib/bbm064.

Tsubaki M, Tomii K, Sese J. 2019. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*. vol 35(2): 309–318. doi: https://doi.org/10.1093/bioinformatics/bty535.

Wang Y, Bryant SH, Cheng T, Wang J, Gindulyte A, Shoemaker BA, Thiessen PA, He S, Zhang J. 2017. Pubchem bioassay: 2017 update. *Nucleic Acids Research*. vol 45(1): 955–963. doi: https://doi.org/10.1093/nar/gkw1118.

Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. 2008. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics.* vol 24(13): 232–240. doi: https://doi.org/10.1093/bioinformatics/btn162.

Zhang ZC, Zhang XF, Wu M, Ou-Yang L, Zhao XM, Li XL. 2020. A graph regularized generalized matrix factorization model for predicting links in biomedical bipartite networks. *Bioinformatics*. vol 36(11): 3474–3481. doi: https://doi.org/10.1093/bioinformatics/btaa157.