UTM Computing Proceedings
Innovations in Computing Technology and Applications
Volume: 1 | Year: 2016 | ISBN: 978-967-0194-81-3

Comparative Analysis using Classification Algorithm for Drug Target Interaction

Nur Hidayu MdNoh[1], Zuraini Ali Shah*[2]

Department of Software Engineering, Faculty of Computing, Universiti Teknologi Malaysia,81310 Johor Bahru, Johor, Malaysia

[1]ayufazrin92@gmail.com, [2]aszuraini@utm.my

Abstract

Machine learning methods based on drug and target protein interaction data in bioactivity databases are one of the popular strategies that are being used in order to efficiently finding novel lead compounds in the drug discovery process. In order to predict whether there will be an interaction occur between the drug and the target, most method use identified similar drug and targets available in the database. Different type of machine learning method including classification approached are used. However, which of the approached used will give a high and most accurate result? Hence, the goal of this study is to compare the performance of classifiers which are Support Vector Machine (SVM), Naïve Bayesian Model, Nearest Neighbors Models (k-NN), Random Forest Model and the combine classifiers which are SVNKNN, SVMRandomforest, SVMNaïvebayes and bagging approach in order to predict the drug-target protein interaction based on their performance in accuracy, precision, recall and area under curve (ROC). The preprocessing method were first done in order to provide the classifier with the appropriate drug-target interaction data.

Keywords: classification approach, machine learning, drug target interaction

1.0 Introduction

Drug target can be characterized as a molecular structure that is associated and identified with a disease, and its movement is either inhibited or stimulated by the drugs that are controlled to battle or analyze the said disease (Imming P *et al*., 2006). Various of research related on drug-target protein interaction are published throughout the year. This is because of the advance studies in molecular medicine and the human genome project which provide huge opportunities to the researcher in order to discover new information and discoveries in the drug-target protein interaction.

There are two main categories of computational approaches in predicting the drug-target interaction, which are the docking simulation and machine learning. In the docking simulation, three dimensional (3D) structure or a large sets of drug for inverse docking are required to make the prediction. This approach is always difficult because of the small size

of the 3D structure. In addition to the problem, the docking simulation consume a lot of time. In contrast, the machine learning approach is more preferable in drug-target interaction prediction as it is more efficient to be used when dealing with a large number of drugs and targets proteins.

Different type of machine learning method including classification approached are used. However, which of the approached used will give a high and most accurate result? Hence, the goal of this study is to compare the performance of classifiers which are Support Vector Machine (SVM), Naïve Bayesian Model, Nearest Neighbors Models (k-NN), Random Forest Model and bagging approach in order to predict the drug-target protein interaction based on their performance in accuracy, precision, recall and area under curve (ROC).Hence, this study focus on the performance of the classification models used in order to predict the drug-target interaction.

The objectives of this research are to study the different type of classification technique that related to ligand domain, to analyze the implementation of different type of classifiers in order to predict the drug-target protein interaction and lastly to compare the performance between each of the single classifiers and combined classifiers used in predicting drug target protein interaction

2.0 Methodology

The GPCR data were collected from Yamanishi *et al*. (2008) were used as the benchmark for the research while the generated negative sets of drug molecules were retrieved from the data used in Cao *et al*. (2012a). Both of the positive and negative dataset were manually combine in a single comma delimitated (.csv) file in order to easily be called into the R software. Two column were created in the file to representing the target id and drug id. The first 635 rows of the data are the positive set, and the rest of it is the negative set. This set are used as the positive class and negative class for the classifiers. After all the sequences for both protein and drugs were retrieved, the descriptor for both protein and the drugs molecule were calculated. These descriptors were calculated by using the function that is available in the Rcpi package. Descriptor is only for unique drug and target list, hence, a full descriptor matrix need to be generated for the data. In the previous phase, the molecular descriptor and protein descriptor has been generated. Hence, by using both of the generated descriptors the drug-target interaction descriptor is generated. This drug target interaction descriptor data will be implemented into the single classifiers and the combined classifiers.

3.0 Result

Table 1 below shows the overall result gained from each of the classifiers used to predict the drug-target interaction. Tenfold cross validation is used in the classifiers chosen in predicting the drug-target interaction in this study. The criteria for evaluating the prediction performance among these classifiers that are being compared are the accuracy, precision, recall, and Receiver Operating Characteristic (ROC) value.

Table 1: Overall result

Classifier	Accuracy (%)	Precision	Recall	ROC Area
Single Classifiers				
Naïve Bayesian model	71.4961	0.747	0.65	0.754
Nearest Neighbor model (k-NN)	75.5906	0.714	0.854	0.838
Support Vector Machine (SVM) model	80.315	0.787	0.831	0.803
Random Forest model	78.189	0.772	0.8	0.848
Combined Classifiers				
SVMKNN	80.315	0.787	0.831	0.881
SVMRandomForest	80.315	0.787	0.831	0.869
SVMNaïveBayes	79.9213	0.77	0.854	0.843
Bagging	77.7165	0.777	0.794	0.836

Figure 1 below show the performance evaluation by precision, recall, and ROC area value between the single classifiers while Figure 2 show the performance evaluation by the combined classifiers.

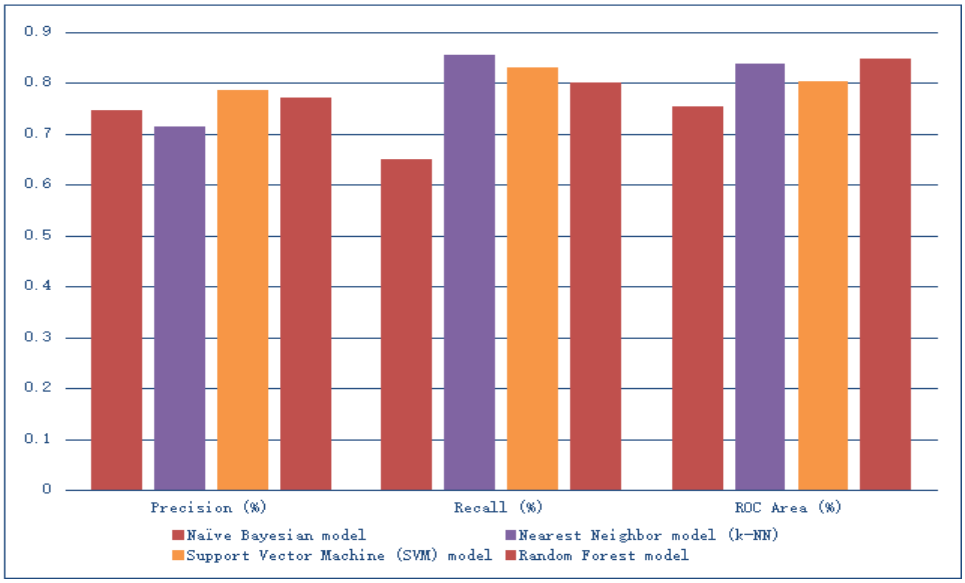

Figure 1: Performance evaluation by precision, recall, and ROC area value between the single classifiers

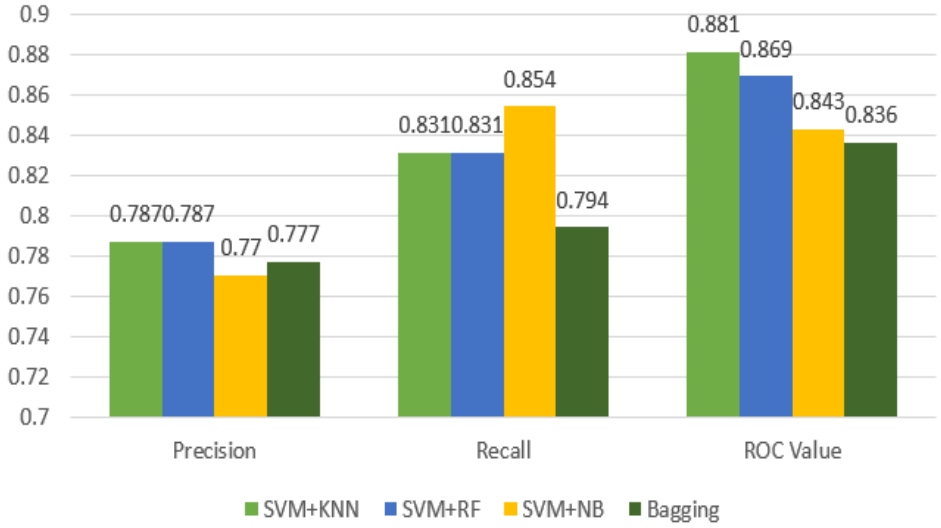

Figure 2: Performance evaluation by precision, recall, and ROC area value between the combined classifiers

4.0 Discussion

There are three objectives of this research which are to study the different type of classification technique that related to ligand domain, secondly is to analyze the implementation of different type of classifiers in order to predict the drug-target protein interaction and lastly is to compare the performance between each of the single classifiers and

combined classifiers used in predicting drug target protein interaction.

Based on the result shown in Table 1, for the single classifiers used, SVM model gain the highest accuracy value in predicting the drug-target interaction which is 80.315%, followed by Random Forest model at 78.189%. The accuracy gain from K-NN model is 75.5906% and last but not least Naïve Bayesian model at 71.4961%. In comparing the precision value of the data, again SVM model have higher value compared to the other classifier. For the recall value, K-NN have a better recall value which is 0.854. Area under curve was represented by the ROC value. Even though SVM model gain the highest accuracy, but it ROC value are lower than Random Forest Model. Higher ROC value denoted that the model can advantageously be applied in virtual screening. From these study SVM and Random Forest appears to be a potentially useful classification tools for prediction of drug-target interaction.

As for the combined classifiers,SVMKNN and SVMRandomForest model gain the highest accuracy value in predicting the drug-target interaction which is 80.315%, followed by SVMNaïveBayes at 79.9213% and Bagging approach at 77.7165%. Meanwhile, in comparing the precision value of the data, again SVMKNN model and SVMRandomForst model have higher value compared to the other classifier. For the recall value, SVMNaïveBayes have a better recall value which is 0.854. Area under curve was represented by the ROC value. SVMKNN give the highest value of area under curve which is 0.881, followed by SVMRandomForest at 0.869, SVMNaïveBayes at 0.843 and Bagging approach at 0.836. Higher ROC value denoted that the model can advantageously be applied in virtual screening. From these study, all these combined classifiers appear to be a potentially useful classification tools for prediction of drug-target interaction.

5.0 Conclusion

Various of research related on drug-target protein interaction are published throughout the year. This is because of the advance studies in molecular medicine and the human genome project which provide huge opportunities to the researcher in order to discover new information and discoveries in the drug-target protein interaction.

In this research, the main goal of the study is to compare the performance between classification model to predict the drug-target interaction by using the accuracy, precision, recall value and ROC under curve. The result show that for the single classifier, SVM give the highest accuracy while random forest gives the highest ROC value which indicate that the model can be used in virtual screening. As for the combined classifiers, SVMKNN and SVM Randomforest give the highest accuracy value. In addition, all the combine classifiers obtained high ROC value and SVMKNN obtained the highest ROV value. This also indicate that all of the combine classifiers used in this research can be used in virtual screening.

References

Cao, D., Liang, Y., Deng, Z., Hu, Q., He, M., & Xu, Q. et al. (2013). Genome-Scale Screening of Drug-Target Associations Relevant to Ki Using a Chemogenomics Approach. *Plos ONE*, *8*(4), e57680. http://dx.doi.org/10.1371/journal.pone.0057680

Cao DS, Liu S, Xu QS, Lu HM, Huang JH, Hu QN, Liang YZ (2012a). "Large-scale prediction of drug-target interactions using protein sequences and drug topological structures."Analytica chimica acta, 752, 1-10.

Sugaya, N. (2013). Training based on ligand efficiency improves prediction of bioactivities of ligands and drug target proteins in a machine learning approach. Journal of chemical information and modeling, 53(10), 2525-2537.

Sugaya, N. (2014). Ligand efficiency-based support vector regression models for predicting bioactivities of ligands to drug target proteins. Journal of chemical information and modeling, 54(10), 2751-2763.

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, 24(13), i232-i240.

Yamanishi, Y., Kotera, M., Kanehisa, M., & Goto, S. (2010). Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatics, 26(12), i246-i254.