INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

An Empirical Study to Analyse The Effect of Bagging and Feature Subspacing on The Performance of A Custom Ensemble Algorithm for Predicting Drug Protein Interactions

Harshita Bhargava[1]*, Amita Sharma[1], Prashanth Suravajhala[2,3]

1 Department of Computer Science & IT, IIS (deemed to be University), Jaipur, Rajasthan, India
2 Amrita School of Biotechnology, Amrita University, Clappana, Kollam, 690525, Kerala, India
3 Bioclues.org, India

OPEN ACCESS

Received: 21-12-2023
Accepted: 30-01-2024
Published: 27-02-2024

Citation: Bhargava H, Sharma A, Suravajhala P (2024) An Empirical Study to Analyse The Effect of Bagging and Feature Subspacing on The Performance of A Custom Ensemble Algorithm for Predicting Drug Protein Interactions. Indian Journal of Science and Technology 17(10): 911-916. https://doi.org/10.17485/IJST/v17i10.3202

* **Corresponding author**.

harshita.bhargava@iisuniv.ac.in

Funding: None

Competing Interests: Noe

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: The objective of this study is to analyse the effect of bagging and feature subspacing on the performance of a custom ensemble of decision tree classifiers for predicting drug protein interactions. **Methods:** In our present work we have designed a custom ensemble algorithm with decision trees as the base learner. We analysed the effect of bagging negative samples and feature subspacing on the performance of the custom ensemble in terms of AUCROC and AUPR. The Enzyme dataset from the Yamanishi dataset composed of 445 drugs and 664 proteins was used for the experiments. **Findings:** It was observed that the effect of bagging negative samples was significant as compared to feature supspacing in terms of AUPR metric. Now since AUPR is a metric that remains unaffected by the presence of negative samples hence the increase in AUPR by increasing the negative to positive ratio clearly indicated that the negative samples do contain the positives which are unknown and are yet to be verified. **Novelty:** The results give a strong indication that that feature subspacing has no considerable impact on the AUCROC metric performance of the custom ensemble while AUPR metric increases as the negative to positive ratio increases. The results give a foundation to the fact that, finding reliable negative samples from the entire set of negative drug protein pairs can further enhance the performance of the machine learning classifiers.

Keywords: Decision tree classifier; Ensemble classifier; Drug discovery; Bagging; Drug repurposing

1 Introduction

In this research work, only feature based methods from the machine learning category of chemogenomic based approach have been considered. Drugs and targets are used in feature-based approaches as pairs, with each drug represented by a feature vector of the

form d=df_1,df_2,...df_n, and each target by t=tf_1,tf_2,...tf_s. The feature vectors of each drug and each target are concatenated to create the pairs.

d⊕t={df_1,df_2,...df_n,tf_1,tf_2,...tf_ss}

The binary labels on these feature vector pairs are "1" to denote an interaction and "0" to denote a non-interaction between the corresponding drug and target pair. If DTI prediction is handled as a regression problem rather than a classification problem, the labels can also be real valued affinity scores (Ki, Kd, IC50, EC50) to indicate interactions and non-interactions. The primary issue with this situation is that the non-interactive pairs are actually pairs for which the interaction is unknown or has not been empirically proved. The majority of methods treat these hypothetical interactions as non-interactions when creating the corresponding models. SVM [1–5], Rotation forest [6,7] and Random forest [8] have been prominently used for DTI prediction problems.

In recent years, feature-based ensemble approaches have attracted a lot of attention as well. The accuracy of a combined classifier/regressor model has been empirically demonstrated to be superior to a single classifier/regressor [9], given that the individual base learners are diverse. Varying the training samples between the learners is one method for infusing diversity which has been termed as "Bagging". The base learners may also be varied to infuse diversity in the ensemble. A custom boosting ensemble CFSBoost was proposed that used feature subspacing on the protein features from different feature categories [10]. In the same sequence [11] addressed the within-class imbalance problem by defining a custom oversampling procedure. The between–class imbalance was addressed using randomly sampling negative instances equal to the number of positive instances for each base learner in the ensemble. The feature subspacing was done on the drug protein pair feature set to ensure diversity in the ensemble. Another ensemble framework was proposed [12] to address the imbalance issue followed by dimensionality reduction on subspaced drug and protein features. The decision tree and kernel ridge regression were used as base learners resulting in two different ensembles EnsemDT and EnsemKRR. In [13] the concept of "Active learning" in addition to dimensionality reduction and feature subspacing were proposed for an ensemble framework based on bagging. Assigning weights to majority and minority samples allowed for the initial sampling of the samples from the training set using neighbourhood balanced bagging. The following phase involved performing feature spacing on the drug and target features independently, followed by dimensionality reduction on the drug and target features, respectively. The research on finding the set of reliable negative samples has gained considerable attention in the recent years [14] [15]. The outcome of this analysis gives an evidence that the set of negative samples contain probable positives which are not known and have not been validated through experiments. Hence, it emphasizes the need to design algorithms with the objective of uncovering the reliable negative samples.

2 Methodology

2.1 Dataset Statistics

The evaluation was done using the gold standard dataset given by Yamanishi et al., 2008. It includes four classes of proteins viz, Enzymes, Ion channel, GPCR and Nuclear Receptor along with the respective interaction data with each of the drugs. It is a classification-based dataset that includes binary 0 and 1 interaction data for each drug and each type of protein. Each dataset under consideration is a discrete dataset made up of various drugs and a distinct class of proteins from the KEGG database. Though in our experiments we used only the Enzyme dataset to gain insights and analyse the effect of feature subspacing and bagging the negative samples on the custom ensemble algorithm. The choice of the dataset was mainly based on its size which contained the maximum number of drugs and proteins for evaluating the model.

The details for the Yamanishi dataset are given below:

Table 1. Statistics for the Yamanishi Dataset

Protein class	# of Drugs	# of Proteins	Known interactions	Unknown interactions	Sparsity ratio
Enzyme	445	664	2926	2,92,554	0.010
Ion Channel	210	204	1476	41,364	0.036
GPCR	223	95	635	20,550	0.031
Nuclear Receptor	54	26	90	1314	0.068

2.2 Data Extraction and Preprocessing

We have utilised the side information of drugs and proteins in the form of their respective features. Initially in order to extract the drug SMILES and protein sequences from the KEGG database using the Rcpi package, an R script was developed. The Mordred tool was used to develop a quick Python script for feature extraction for medications. The 2D category of the molecular descriptors comprised geometrical, topological, and constitutional descriptors. Using the Propy3 tool, the features of proteins were obtained, including the descriptors for composition transition and distribution (CTD), Moran autocorrelation, and amino acid composition (AAC) and dipeptide composition (DPC). The proposed method utilised a total of 1613 drug descriptors and 807 protein descriptors for the Enzyme dataset.

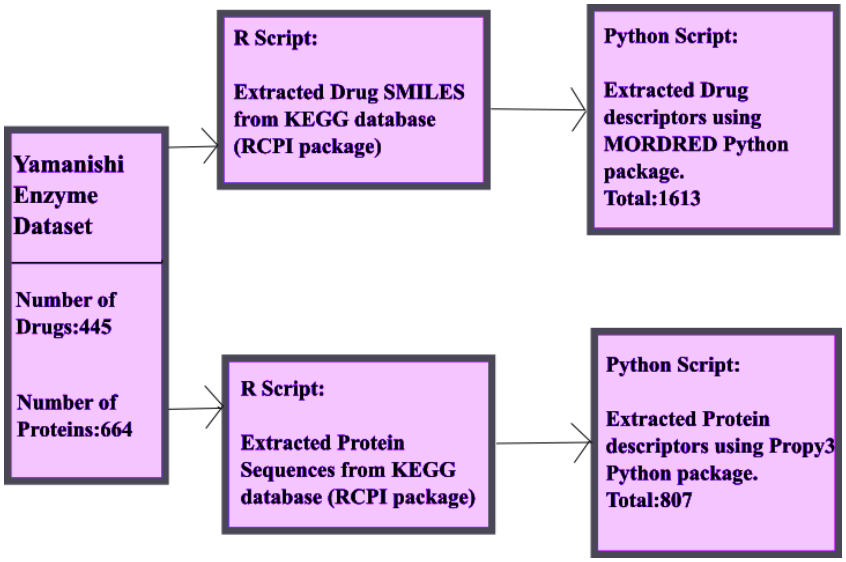

Fig 1. Drug/Protein descriptors extracted using Mordred and Propy3 tool

2.3 Drug descriptors (2D) extracted using Mordred tool

ABCIndex, AcidBase, AdjacencyMatrix, Aromatic, AtomCount, Autocorrelation, BalabanJ, BaryszMatrix, BCUT, BertzCT, BondCount, CarbonTypes, Chi, Constitutional, DetourMatrix, DistanceMatrix EccentricConnectivityIndex, EState, ExtendedTopochemicalAtom, FragmentComplexity, Framework, GravitationalIndex, HydrogenBond, InformationContent, KappaShapeIndex, Lipinski, McGowanVolume, MoeType, MolecularDistanceEdge, MolecularId, PathCount, Polarizability, RingCount, RotatableBond, SLogP, TopologicalCharge, TopologicalIndex, TopoPSA, VdwVolumeABC, VertexAdjacencyInformation, WalkCount, Weight, WienerIndex, ZagrebIndex.

2.4 Protein descriptors extracted using Propy3 tool

- Dipeptide composition descriptors.
- Amino acid composition descriptors.
- Composition transition and distribution descriptors.
- Moran autocorrelation descriptors.

Drug (or protein) features with constant values or unique values less than two in a single column were eliminated from the drug (or protein) feature set on the part of preprocessing. To deal with non-numeric values, the features with NAN values were changed and filled with 0. The drug-protein pair feature vectors were created by concatenating the drug feature vector (Df_1,Df_2,.....Df_n) and protein feature vector (Pf_1,Pf_2,....Pf_m) in each of the enzyme, ion channel, GPCR and nuclear receptor datsets. As a result, the feature vector for the drug-protein pair was:

DP_{pair} = (Df_1,Df_2,.....Df_n ,Pf_1,Pf_2,....Pf_m)

where n indicates the total number of drug features and m indicates the total number of protein features.Thus, a total of 2,95,480 (445*664) drug-protein pairs were considered for the enzyme dataset for evaluating the proposed method. On a similar

note, there were 210*204 = 42,840,223*95 = 21,185,54*26 = 1404 total drug-protein pairs for Ion channel, GPCR and Nuclear receptor datasets respectively.

3 Results and Discussion

3.1 Experimental Design

Python 3

- Google colab pro+
  - RAM:50GB
  - Disk space:250GB

3.2 Experiment 1.1

Based on the EnsemDT algorithm developed by Ezzat et al. (2017), we created a ”custom ensemble” of decision tree classifiers for this experiment. The ensemble was built using 20 base learners, and to add variation/diversity, feature subspacing on the protein feature set and drug feature set independently was done for each base learner. We used ”s” as the subspacing parameter and examined the protein and drug feature sets with values of 0.1, 0.2, 0.3, 0.4, 0.5, and 0.6 respectively. Bagging was done on negative samples so that, when building each base learner, the negative to positive ratio varied as 5, 10, and 15 respectively. This ratio indicated that there were either 5, 10, or 15 times as many negatives as positives. The major goal of the experiment was to ascertain how the custom ensemble's performance would be affected by feature subspacing and bagging negative samples. We used AUCROC and AUPR as the performance metrics, where AUPR focuses on minority positive samples and remains unaffected by the presence of negative samples in the dataset.

Table 2. Performance analysis of custom ensemble using subspacing and bagging on negative samples with negative to positive ratio 5:1 (Enzyme dataset)

Subspacing parameters	Bagging on negatives (Number of negatives=5 times the number of positives)	
	AUCROC	AUPR
s=0.1	0.86	0.29
s=0.2	0.86	0.3
s=0.3	0.85	0.3
s=0.4	0.87	0.3
s=0.5	0.86	0.28
s=0.6	0.87	0.31

Table 3. Performance analysis of custom ensemble using subspacing and bagging on negative samples with negative to positive ratio 10:1 (Enzyme dataset)

Subspacing parameters	Bagging on negatives(Number of negatives=10 times the number of positives)	
	AUCROC	AUPR
s=0.1	0.81	0.43
s=0.2	0.83	0.44
s=0.3	0.83	0.46
s=0.4	0.83	0.41
s=0.5	0.83	0.47
s=0.6	0.83	0.43

Table 4. Performance analysis of custom ensemble using subspacing and bagging on negative samples with negative to positive ratio 15:1 (Enzyme dataset)

Continued on next page

Table 4 continued

Subspacing parameters	Bagging on negatives (Number of negatives=15 times the number of positives)	
	AUCROC	AUPR
s=0.1	0.80	0.50
s=0.2	0.80	0.49
s=0.3	0.80	0.47
s=0.4	0.81	0.47
s=0.5	0.81	0.49
s=0.6	0.81	0.46

- **Inference**

Tables 2, 3 and 4 make it clear that the feature subspacing has no discernible impact on the AUCROC metric performance of custom ensemble. The only exception, when the ratio of negative to positive data is 5:1, this statistic rises as feature subspacing increases.

As opposed to the previous two cases, the growth in this instance was not gradual. The performance declined in terms of the AUCROC metric but improved in terms of the AUPR metric when the negative to positive ratio increased from 5 to 15. Although the rise in AUPR performance metric, with subspacing on drug and protein features was not found to be steady.

3.3 Experiment 1.2

In this experiment, the custom ensemble was built using 20 base learners, bagging the negative samples, and not using feature subpacing as in experiment 1.1. The negative to positive ratio was adjusted to be 5, 10, and 15 for each base learner, accordingly.

The main objective of the experiment was to examine the effect of bagging the negative samples on ensemble's performance.

Table 5. Performance analysis of custom ensemble while bagging on negative samples for each base learner with different negative to positive ratio (Enzyme dataset)

Bagging on Negative samples	AUCROC	AUPR
Number of negatives=5 times the positives	0.86	0.28
Number of negatives=10 times the positives	0.83	0.42
Number of negatives=15 times the positives	0.81	0.47

- **Inference**

Table 5 makes it very evident that, while bagging occurs on negative samples, the AUPR rises as the negative to positive ratio rises. Due to AUPR's exclusive focus on the minority class, the improvement in performance for various negative to positive ratios suggests that the negative samples really contain true positives that were mistakenly labelled as negatives. However, these are actually unknowns that have an impact on the classifier's performance.

4 Conclusion

We built a custom ensemble using decision trees as the base learner to observe the effect of bagging on negative samples and feature subspacing parameter. The experiments revealed that feature subspacing has no considerable impact on the AUCROC metric performance of the custom ensemble while AUPR metric increases as the negative to positive ratio increases. Since AUPR primarily deals with the minority class hence increasing the negative to positive ratio clearly indicates that the negative samples actually contain true positives which have been taken as negatives. In order to build reliable supervised machine learning (ML) models, there is a need to find reliable negative samples.

References

1) Nagamine N, Sakakibara Y. Statistical prediction of protein–chemical interactions based on chemical structure and mass spectrometry data. *Bioinformatics*. 2007;23(15):2004–2012. Available from: https://doi.org/10.1093/bioinformatics/btm266.
2) Chen R, Liu X, Jin S, Lin JS, Liu J. Machine Learning for Drug-Target Interaction Prediction. *Molecules*. 2018;23(9):1–15. Available from: https://doi.org/10.3390/molecules23092208.

3) Faulon JLL, Misra M, Martin S, Sale K, Sapra R. Genome scale enzyme–metabolite and drug–target interaction predictions using the signature molecular descriptor. *Bioinformatics*. 2008;24(2):225–233. Available from: https://doi.org/10.1093/bioinformatics/btm580.
4) Jacob L, Vert JP. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics*. 2008;24(19):2149–2156. Available from: https://doi.org/10.1093/bioinformatics/btn409.
5) Tabei Y, Pauwels E, Stoven V, Takemoto K, Yamanishi Y. Identification of chemogenomic features from drug–target interaction networks using interpretable classifiers. *Bioinformatics*. 2012;28(18):i487–i494. Available from: https://doi.org/10.1093/bioinformatics/bts412.
6) Rodriguez JJ, Kuncheva LI, Alonso CJ. Rotation Forest: A New Classifier Ensemble Method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2006;28(10):1619–1630. Available from: https://doi.org/10.1109/TPAMI.2006.211.
7) Wang L, You ZH, Chen X, Yan X, Liu G, Zhang W. RFDT: A Rotation Forest-based Predictor for Predicting Drug-Target Interactions Using Drug Structure and Protein Sequence Information. *Current Protein & Peptide Science*. 2018;19(5):445–454. Available from: https://doi.org/10.2174/1389203718666161114111656.
8) Yu H, Chen J, Xu X, Li Y, Zhao H, Fang Y, et al. A Systematic Prediction of Multiple Drug-Target Interactions from Chemical, Genomic, and Pharmacological Data. *PLoS ONE*. 2012;7(5):1–14. Available from: https://doi.org/10.1371/journal.pone.0037608.
9) Hansen LK, Salamon P. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1990;12(10):993–1001. Available from: https://doi.org/10.1109/34.58871.
10) Rayhan F, Ahmed S, Farid DM, Dehzangi A, Shatabda S. CFSBoost: Cumulative feature subspace boosting for drug-target interaction prediction. *Journal of Theoretical Biology*. 2019;464:1–8. Available from: https://doi.org/10.1016/j.jtbi.2018.12.024.
11) Ezzat A, Wu M, Li XL, Kwoh CK. Drug-target interaction prediction via class imbalance-aware ensemble learning. *BMC Bioinformatics*. 2016;17(S19):267–276. Available from: https://doi.org/10.1186/s12859-016-1377-y.
12) Ezzat A, Wu M, Li XL, Kwoh CK. Drug-target interaction prediction using ensemble learning and dimensionality reduction. *Methods*. 2017;129:81–88. Available from: https://doi.org/10.1016/j.ymeth.2017.05.016.
13) Sharma A, Rani R. BE-DTI': Ensemble framework for drug target interaction prediction using dimensionality reduction and active learning. *Computer Methods and Programs in Biomedicine*. 2018;165:151–162. Available from: https://doi.org/10.1016/j.cmpb.2018.08.011.
14) Najm M, Azencott CAA, Playe B, Stoven V. Drug Target Identification with Machine Learning: How to Choose Negative Examples. *International Journal of Molecular Sciences*. 2021;22(10):1–15. Available from: https://doi.org/10.3390/ijms22105118.
15) Sharifabad MM, Sheikhpour R, Gharaghani S. Drug-target interaction prediction using reliable negative samples and effective feature selection methods. *Journal of Pharmacological and Toxicological Methods*. 2022;116:107191. Available from: https://doi.org/10.1016/j.vascn.2022.107191.