iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VI **Month of publication:** June 2021

DOI: https://doi.org/10.22214/ijraset.2021.35716

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

A Deep Learning Approach for Drug Target Interaction Prediction

Uday Verma[1], Kunal Gupta[2], Ishita Goyal[3]
[1, 2, 3]*Student (CSE), Maharaja Agrasen Institute of Technology, Delhi*

Abstract: ***Computer simulation so called 'in silico' techniques for prediction of drug target interactions constitute as a critical phase in the process of efficient, cost effective and reliable development process. Drug-Target interaction (DTI) plays an important role in drug discovery, drug repositioning and understanding the side effects of the drugs which helps to identify new therapeutic profiles for various diseases. Therefore, developing computation method to predict possible Durg-Target combinations-interactions with less probability of High positive rates is necessary. Here we incorporate Deep Learning approach with graph based computational method for Drug Target Interaction Prediction. We combine similarity based as well as feature selection-based methods with exploiting Graph techniques such as Graph embedding, graph mining so as create a model based on the heterogenous network. Heterogenous network is thus constructed by supplementing the known drug-target interaction graph with drug-drug and target-target similarities graph in order to draw terminal heterogenous graph after using similarity selection method procedure and algorithm. Compared to other computer methods developed to predict DTI, we achieved a significantly improved prediction score using four benchmark sets of data. AUPR score across all databases (0.92) which is improved by over 30% relative to second best performing model.***

Keywords: Drug-Target interaction, Machine Learning, drug discovery, Adaboost, Graph embedding, Neural Network

I. INTRODUCTION

The high cost, low success rate, and tediousness of conventional experimental drug discovery processes have led to the introduction of cheap computer methods that can facilitate drug discovery and development. [1]. With this in mind, computer-aided methods for predicting drug target interactions (DTI) are being sought to narrow the research field to more sustainable drugs. The first step in understanding which drugs to use depends on their ability to interact with specific target proteins to enhance or inhibit their function. However, the number of DTI pairs determined and verified through experiments may be limited. Thus, predicting DTI is a necessary task within Early evaluation of potential new drugs and finding new uses for existing drugs; that is, drug reuse. To date, several different approaches are used for predicting DTIs, they all have limitations and require significant improvements in predictive properties.

Two important computer methods are docking modeling and machine learning. Combination modeling is very common in biology, but it has two serious problems: (1) We need to know the three-dimensional structure of the target molecule in order to calculate the binding of each drug candidate to the target molecule [2-4], but many three-dimensional structure of the target, especially the GPCR, is not yet available; (2) The simulation is very time-consuming because it requires a lot of computing resources. whereas, machine learning is effective and more efficient than docking simulation in making larger predictions And then choose more promising candidates for more experimental choices. Here, we emphasis on deep learning methods for predicting DTIs. The methods use three kinds of information: drug-specific information (such as chemical information about the drug), target-specific information (such as protein sequence), or known information from DTI. It is mainly divided into three categories, termed as: machine learning (ML)-based methods [5-7], and deep learning (DL)-based methods. (DL is a branch of ML) and network methods.

The prediction of drug-target interactions is currently receiving a lot of attention in bioinformatics and chemoinformatics. More in-depth comparison and analysis of ML algorithms will help biologists and chemists choose the most appropriate model, and computer scientists are developing prediction methods with higher throughput. Similarity-based methods require closer inspection. More information on further improving prediction performance can be found hereFor DTI, we propose a calculation method that uses topological information and the similarity of multiple drugs and targets. This method of using graph embedding, graph mining, and similarity-based techniques to predict drug-target interactions is applied to the prediction problem of predicting DTI as heterogeneous graph relationships.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

Table 1 Optimum Yamanishi_08 dataset stats

Statistics	NR	GPCR	IC	Enzyme
No. of drugs	54	223	210	445
No. of targets	26	95	204	664
Known DTIs	90	635	1476	2926
Unknown DTIs	1314	20,550	41,364	292,554
Sparsity ratio	0.068	0.031	0.036	0.010

II. LITERATURE SURVEY

Machine learning methods have been developed using feature-based methods, where feature vectors represent DTI, and Similarity-based methods based on the "guilty association" principle [8]. Some of the early work by Yamanishi and his co-authors to successfully predict DTI based on controlled ML was the use of pharmacological, chemical, and genomic data. Various methods based on these assumptions are summarized in [8], most of which give promising results. Formulate DTI prediction as a channel prediction problem on a non-uniform graph. For example, DASPfnd [9] uses a drug-drug similarity matrix and a target-target similarity matrix to draw DTI, and then DASPfnd sorts the DTIs according to the DTI single pathway score to find the 1% higher DTI. This method outperforms several network methods when predicting the highest-rated individual using the DTI reference data set to view, Yamanishi_08 [10]. Since all drug-to-drug (or target-to-target) similarities and DTI can be expressed as adjacency matrices, matrix factorization methods have recently been integrated into ML and/or network methods. Based on the method of DTI prediction. The graph inclusion technology applied to the knowledge graph also improves the DTI prediction performance through training. An introduction to the low-dimensional properties of the drug or target used in the ML or DL method. For example, DTINet [11] uses matrix factorization and graph embedding methods to predict new DTI from heterogeneous graphs. DTINet gathers information on different types of drugs and target proteins, including drug-disease associations, drug-effect associations, drug similarities, drug interactions, protein-protein interactions, and protein-disease.Construct a complete heterogeneous map of protein-protein associations and similarities. DTINet uses matrix factorization to construct the objective function, and then learns to capture the low-dimensional feature representation of the topological properties of each node in this heterogeneous graph. This method is superior to other modern methods using HPRD and DrugBank data sets. However, DTINet does not effectively predict drug interactions or target interactions, which is considered a shortcoming of this method. [11].

III. APPROACH

A. Datasets

Machine learning algorithms work on data provided to the defined ml model. Data on drugs, targets, and drug-target interactions are available from the following databases: here we use four collected and compiled standard data sets (Yamanishi_08) [10], which are usually used to evaluate newly developed DTI, method for assessing the quality of predictions. Each of the four datasets, namely Enzyme (E) constituting the relation of drug and enzyme, ion channel (IC) with relation among drug and ion channel, G protein coupled receptor (GPCR), and Nuclear receptor (NR), representing relations with drug and respected target, are among the remarkable group of protein targets.

Table 1 provides statistics for all data sets used in this study. The variance index is the number of known DTI divided by the number of unknown DTI, reflecting the imbalance between positive and negative samples (see Table 1).

B. Preprocessing and Similarity Calculation

Based on the principle of guilt-association, similar drugs can have similar targets, and vice versa, as shown in Figure 1. In our method, we use different similarity measures (ie kernels) in each drug pair or between target pairs (ie proteins). In order to collect information from different sources from different views, different drug-drug similarity and target-target similarity are calculated.

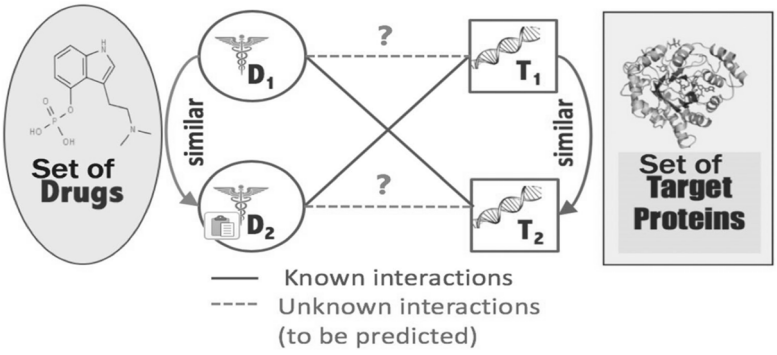


Fig. 1DTIs prediction problem depiction

C. Multiple Similarity Between Drugs Pairs

According to research [12], we calculated or extracted 10 representations or features that determines drug-drug similarity. Using the distinct six views of the similarity of drugs with regard to chemical structure (MOL, SDF, or SMILES format), including SIMCOMP similarity [10] and spectral similarity matrix, marginalization, Lambda-k, Tanimoto and Min-Max-Tanimoto are calculated based on Rchemcpp, KEGGREST[14] and ChemmineR. In the same way, three different distinct views in the study [10], determine drugs similarity based on the side effects, considering SIDER similarity matrix, AERS frequency, and AERS bit. The tenth drug similarity is calculated based on the Gaussian interaction curve (GIP) introduced in, which projects the structure of the drug target network as a network interaction curve. Array with name and source.

D. Multiple Drug target- drug Target Similarity

Similar to drug similarity, we calculated or extracted 10 similarity matrices for the research target [10]. According to the amino acid sequence of the protein, seven different representations reflect the similarity between the targets, including a normalized Smith-Waterman (SW) score [13] and two spectral similarity matrices (k metric equals 3 and 4) . Similarity matrix (different parameters with k length measures andThe maximum number of mismatches for each k metric), recalculated using the R package, KEGGREST [14] and KeBABS [15]. Use GO.db package and R annotation to calculate the genetic ontology similarity matrix (GO) based on GO conditions. As a result of the research, the similarity of protein-protein interaction (PPI) is obtained, which reflects the shortest distance between any pair of targets in the PPI network. GIP is calculated for targets, just like we calculate for drugs. Table 1 shows the target similarity matrices with their names and sources.

IV. METHODS

A. Problem

In this, we pursue a network-based approach. We have defined a weighted heterogeneous graph that is represented by the network of DTIs, supplemented by the drug-drug similarity graph and the target-target similarity graph. This defined graph G (V, E) consists of a set of drugs D = {d1, d2, ..., dn} of n drug nodes and a set of goals T = {t1, t2, ...,tm} of m destination nodes. The DTI G diagram contains three types of edges. The first type of boundary represents the interaction between the agent and the target nodes, and boundaries of this type have been assigned a weight of 1. The two and three threshold types constitutes the similarity among the drugs and the similarity among the targets. These threshold varieties are weighted, with the actual value being between 0 and 1 (0.1]. Given G-Graph, we define the DTI prediction problem as a problem to find links or relation, whose purpose is to predict the unknown correct interaction (represented by the association) between the drug and the target molecule (see Figure 1).

We make various pairings between drugs and targets. Create a "negative sample". Creating this "negative pattern" involves creating connections (ie unknown interactions) between the drug and the target site without boundaries. Therefore, as with other existing calculation methods, we use a set of consistent DTIs as positive interactions, and use randomly generated drug-target pairs to generate negative DTIs.

Therefore, the existing drug-target interaction in the positive pool is removed from the randomly generated drug-target pair to generate a negative interaction. This is because for most drug pools and target groups, there are not enough experiments to prove negative interactions. In our work, we believe that random matching is quite good for negative interaction, because the ratio of current (positive) to non-existent (unknown, negative) k DTI is very small. Then we use different methods to extract the characteristics of each target drug pair.The feature vector is expressed as X = {x1, x2, ..., xn * m} and its label Y = {y1, y2, ..., yn * m}, where n * m is the number of drugs multiplied by the target The number, which is the number of all possible pairs (drugs, targets). If the interaction of the target drug pair is known, the y class of the pair is designated as 1 (y = 1); otherwise the class label is empty (y = 0). Therefore, this is a binary classification task. The goal is to find a new DTI with high accuracy and low false alarm rate. The proposed method combines various techniques related to similarity, features, and graph-based machine learning methods for DTI prediction.

B. Similarity-Based Algorithm

1) *Similarity Integration Method:* We use different integration functions to combine several similarity matrix network synthesis algorithms (SNF). The similarity feature is denoted by a square matrix, see figure 2. The method shown in Figure two iteratively merges these networks and updates each of them Use the nearest neighbor K (KNN) network with information from other networks (this makes the similarity criterion more biased at every step). If networks merge into one network after multiple iterations, SNF stops. Some more information about the SNF function and its parameters are given in [16].
2) *Similarity Selection Method:* How to choose the optimal subset in the similar selection processRobust and should improve the prediction problem, we apply direct similarity selection- Feature Subset Selection(FSS) as a process to obtain the most optimum combination of similarities. FSS follows the same concept as direct feature selection, in which active ingredient-active ingredient and target-target-similarity pairs are added "greedily" until performance improves. More precisely, the records of FSS-the algorithm is a list of all drug similarity matrices (all_DDsim) and a list of all target-to-target similarity matrices (all_TTsim). The algorithm initializes two further lists: an empty list (DDsim) for adding the selected drug and drug similarity matrix, and another empty list (TTsim) for adding the selected target-target similarity matrix. FSS starts with drug-to-drug similarity and target-to-target similarity, and iteratively delivers it, List the combinations: all_DDsim and all_TTsim, and report the results of all these combinations. Select the drug-drug and target-target similarity pair with the best results in DDsim and TTsim as the first set of similarities. In the second round, we have a solid drug-drug similarity and a target-target similarity, we add another unique similarity to the drug and target-target lists and combine them using SNF with a report of all results. Similarities with better results have also been added and corrected in DDsim and TTsim. We repeat these steps, adding and merging the best-performing similarities into the selected similarity set in each round, and stop repeating them only when the results converge (that is, they have not improved).The result is used to generate the G1 graph. Graph embedding for studying features For a given graph G = {V, E}, the graph embedding method converts the graph G to Rd, where d≪ | v |, a node in the graph, its feature vector is much smaller than the graph The actual number of nodes while maintaining the structure and properties of the graph [17].To this end, we use the algorithm structure node2vec to apply training to represent the features in the complete heterogeneous graph G, which consists of the training part of the known DTI, the DT edge in the test data, the drug-similarity Matrix (DD-Sim) and target similarity (TT-Sim) (Figure 4). In order to shorten the processing time of node2vec, we have removed weak edges that do not provide information, active pharmaceutical ingredients, and targets. That is, for each drug (or target), we keep the first k drugs (or targets) similar and delete all other boundaries. After all the weaknesses are eliminated, the similar graphs of drug-drug and target-target KNN are added to the DTI training part and transferred to the node2vec model. After node2vec is applied to the heterogeneous G graph to determine the characteristic representation of each drug and target, Calculate the similarity(cosine) between each drug pair and each target pair to represent two new matrices. These matrices are, Md, an n * n drug similarity matrix, where n is the number of drugs, and Mt, an m * m target-to-target similarity matrix, where m is the number of drugs target; they are used to create G2 graphs. After calculating the cosine similarity between drug pairs (or targets), new edges based on structural and topological similarity may appear. These edges are not similar to the KNN drug in the main image. As well as the similarity of KNN targets, further avoiding the loss of important information.

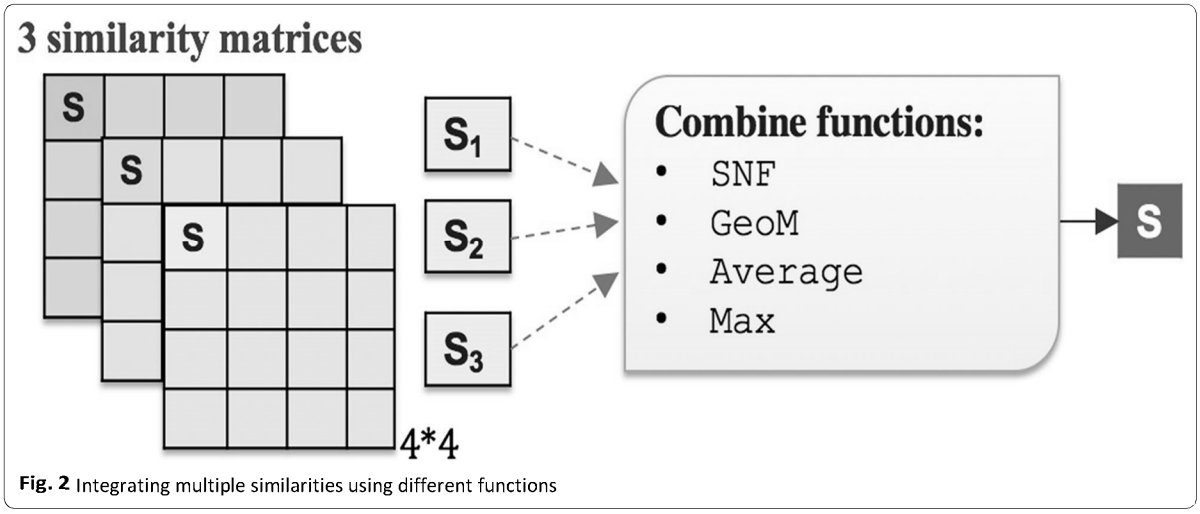


Fig. 2 Integrating multiple similarities using different functions

In order to use and maintain the best hyperparameter set from node2vec, a grid search algorithm is applied to the validation data.

C. Graph-Based Feature Extraction

It is used to evaluate drug target path-score. In this phase, the two graphs G1 and G2 having dissimilar weights are used for extracting features from the graphs. The path evaluation is calculated from the source of each simple path. The node (i. Drug) ends with the target node (i. target protein) of each drug-target pair, using path estimation similar to the introduced DASPfind path score [9], using the following formula:

$$score(d_i, t_j) = \sum_{p=1}^{n} \prod (P_{weights}) \quad \text{--------- (1)}$$

Where P = {p1, p2,..., pn} is a set of paths connected to target j by drug i. In our study, in order to limit or decrease the computational workload the path length is set to be less than or equal to 3 (thus the path length is 2 or 3). This concludes that, there are six potential path structures that exists and which are Ch = {C1, C2, C3, C4, C5, C6} (and are termed in [18,19] as path categories); For instance each starts with a drug node and a destination Target node, and each node along the path appears only once (no loop).The six path structures include two path structures with path length = 2 (C1: (D-D-T) and C2: (D-T-T)) and four path structures with length = 3 (C3: (DDDT) , C4: (DTTT), C5: (DDTT) and C6: (DTDT)). We calculate and determine two characteristics of each routing structure: 1/sum of all meta-route estimates for each routing structure and 2/maximum score of all meta-path score for each routing structure under each path structure. They have an equivalent path structure, and also the meta-path score is that the product of the weights of all edges within the path structure from the drug begin node to the ultimate target node. Rijh represents a series of realations between drug pairs, and the objective equations used to characterize the structure of each path are defined to ensure that the longest path in our method is not discriminated (maximum or summation)Independently, each estimate considers all path sets belonging to a particular path structure. This means that estimates from different route structures will not be mixed into one element. In addition, the score is further normalized using min-max normalization to ensure that the classifier treats the features the same., G1 and G2) (refer to DTI prediction selection model), combinely forming 24-dimensional feature vector.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

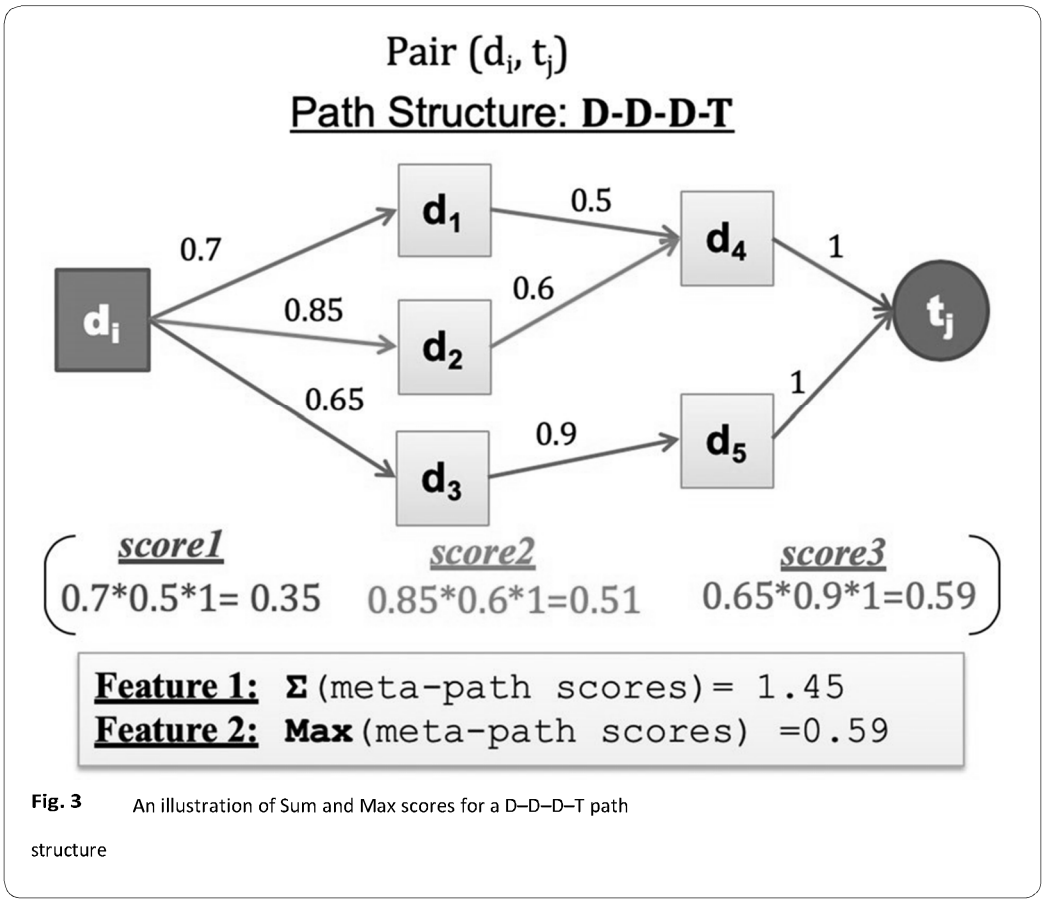


Fig. 3 An illustration of Sum and Max scores for a D–D–D–T path structure

Figure 3 shows an example to illustrate the graph-based feature extraction process through the path structure D-D-D-T. To speed up the execution time, we obtain the path score estimate by applying 3D matrix multiplication. We use an adjacency matrix to represent each graph, which includes the drug-drug adjacency matrix (DD_sim), the target adjacency matrix (TT_sim), and the drug-target interaction matrix (DTI).The path score estimation of each path structure is represented by the matrix multiplication operation. The length of each path structure is equal to the number of products of the adjacency matrix. If the path length = 3, for example D-T-T-T, multiply the 3 matrices to get the same result. For the estimated sum attribute, it is sufficient to perform the usual matrix multiplication, and the resulting matrix represents the sum attribute.However, for the maximum score function, a 3-D matrix multiplication is performed to obtain the multiplied value of each path structure (ie the multiplied edge score), and then the maximum score is selected instead of the summation process. The multiplication of the matrix corresponding to each path structure, and the semantic meaning of each path structure.

D. DTIs Prediction Model

1) *Feature Selection:* The accuracy of the predictive model is based on the identification of the main and necessary features of the researched data set. Therefore, empirical analysis and many experiments (using a concept similar to the direct-forward feature selection method) have been carried out to determine the most important set of functions for this classification problem. Performance analysis includes deleting one or more functions.Therefore, after applying the feature selection step, the dimensionality of the feature vector input to the predictive model is reduced from 24 to the range of 18 to 20 features, concerned with data. Pertaining to the number of known DTIs, as shown in Table 1, we apply oversampling technique or method to the training data, to reconcile data. Act is random oversampling or synthetic minority. Oversampling is applied on the known positive DTI (minority class) such that the training data contains the same number as the main class having negative unknown DTIs. Both methods are implemented using the Python imbalance learning package.

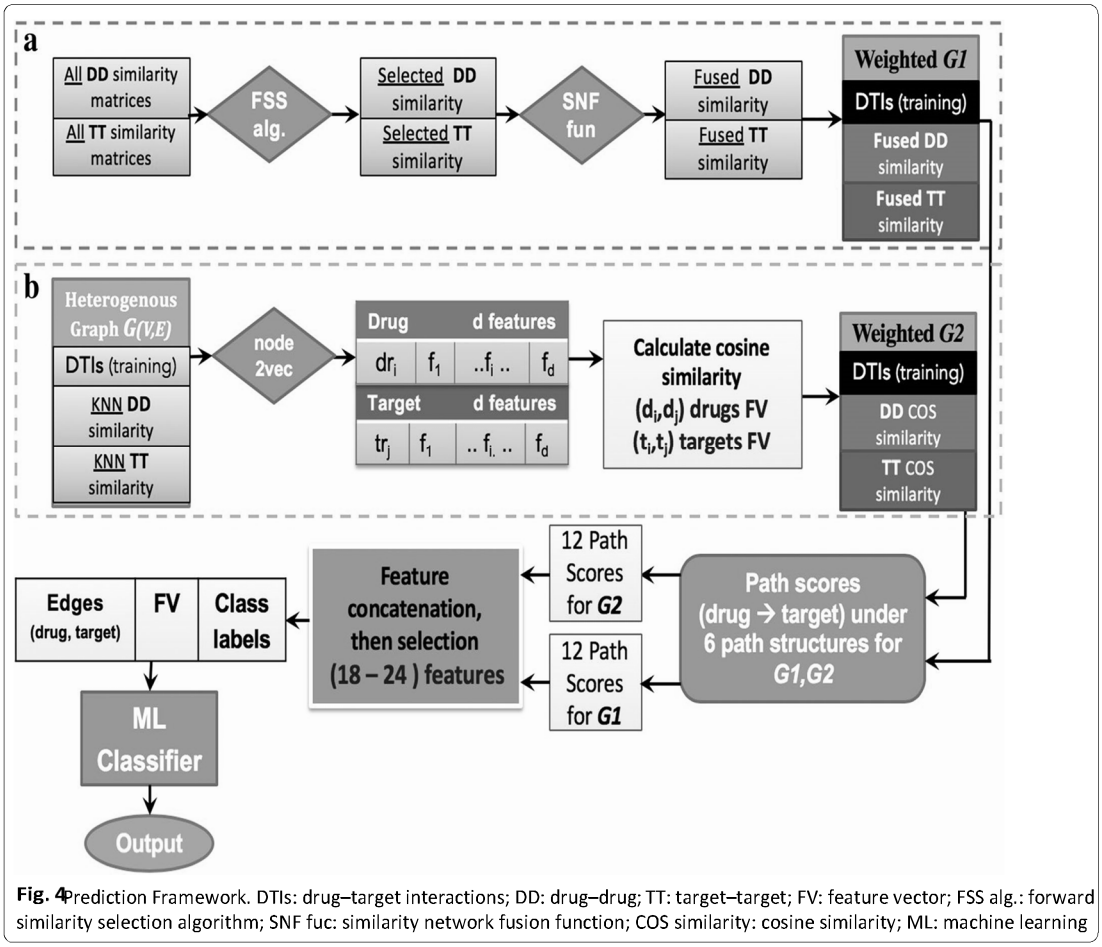


Fig. 4 Prediction Framework. DTIs: drug–target interactions; DD: drug–drug; TT: target–target; FV: feature vector; FSS alg.: forward similarity selection algorithm; SNF fuc: similarity network fusion function; COS similarity: cosine similarity; ML: machine learning

Random oversampling helps to improve the sorting performance of some data sets, while Use the training data set to optimize the parameters to improve the performance of the classifier. For instance the various parameters of the NN classifier are the batch size, the trigger function, the size of hidden nodes and layers, and, while the parameters of the RF classifier include the number of trees, the maximum tree depth, and the number of features. When looking for the best segmentation, consider functions such as measuring the quality of separation.On the other hand, we use Adaboost to control the decision tree classifier to optimize parameters similar to those used in RF. The input data of these classifiers are the feature vectors X and Y labels of all possible active ingredient drug-target pairs.

2) *The Framework:* Feature Vector X is derived using step-wise structure and framework (as shown in Figure 4) for all drug-target pairs used to predict missing edges (unknown DTI is a positive interaction). Two graphs (G1 and G2) are used to generate X. We generate graph G1 as follows: (1a) apply the FSS method to all similarities DD and TT to select the best subset of similarities, (2a) integrate these selected similarities with the SNF algorithm, and then Twelve path scores were extracted for each graph from the six path structures. Then apply feature selection in steps (5) and (6) to eliminate weak features, and then generate a feature vector X = {x1, x2, ..., xn * m} whose label Y = {y1, y2 ,... , yn * m} is used for all active ingredient-target pairs input to the monitored machine learning prediction model using NN, RF or Adaboost classifiers. The classifier is a category label, which is a positive label or a negative label.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

E. Evaluating Techniques

1) Evaluation Matrix: The area under the Receiver Performance Curve (ROC) (AUC) and the area under the Accuracy Recovery Curve (AUPR)[20] are calculated from test data to evaluate the prediction accuracy. The evaluation involvoes calculation of AUC and AUPR[21]. For this we have calculated False Positive Rate(FPR), accuracy or precision[1], recall rate(which is also known as true positive rate or sensitivity based upon false positive (FP), true negative (TN), true positive (TP) and false negative (FN). The values are shown in formulas 2, 3 and 4, respectively.

$$FPR = \frac{FP}{TN + FP} \quad (2)$$

$$Recall = TPR = \frac{TP}{TP + FN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

By these recall vales and FPR values ROC curve is generated for threshold values and AUC is calculated. AUPR is estimated based on the different precision and recall values at each intersection cut-off points, used to draw the curve, and then the area under the curve is estimated. The closer the AUC and AUPR values are to 1, the better the performance. For very unbalanced (i.e. the number of unknown DTI is much higher compared with known DTI data), AUC is seen as an overly optimistic score indicator for predicting DTI, and in this case, AUPR is seen as a better estimate of imbalanced data as it helps to separate predicted score of known and unknown interactions. Scoring metrics are compared with more advanced methods, but we also calculated the error rate (ER) and relative error rate reduction of the best performance model and the sub-optimal performance model (ΔER) defined in the equation.5 or 6:

$$ER = 1 - AUPR \quad (5)$$

$$\Delta ER = \frac{(ER_2 - ER_1)}{ER_2} \quad (6)$$

Table 3 Average scores: AUPR, AUC, and rank position for comparison methods

Methods	BLM-NII	KronRLS	RLS-WNN	NRLMF	DNILMF	DDR	TriModel	Proposed
Avg AUPR	0.68	0.73	0.76	0.80	0.78	0.87	*0.88*	*0.92*
Avg AUC	0.92	0.90	0.96	0.95	0.95	0.96	*0.98*	*0.99*
Avg of the ranking position across all datasets	8	7	6	4	5	3	*2*	*1*

We rounded-off all results to two decimal places. The italic font with underline indicates the best result in each category, while the italic font only indicates the second-best result

V. RESULT

We compared the performance of our proposed method for DTI prediction with other methoda, and used different databases to test the newly predicted DTI. We also highlighted some potential features that may affect prediction performance.

A. *DTI Prediction Performance*

To analyse our method, we compared the prediction performance of DTI with seven current methods using the Yamanishi_08 reference data set. Modern methods include TriModel, DDR [24], DNMLF, NRLMF [23], KronRLS-MKL [22], RLS-WNN and BLM-NII). ...We chose these methods to compare the performance of ours DTI prediction method and the performance of network-based prediction (Graph-based) and/or matrix factorization methods, as they are all incorporate similarity methods along with machine learning, while using previous information to combine and integrate multiple similarity meathods from different sources. Ten times the random CV settings, scores and optimal parameters provided by each method, Our method is superior to all other methods. We achieve the best performance in all data sets (highest average AUPR score = 0.92 and highest average AUC score = 0.99), using the sub-optimal method (TriModel) is 4% higher than the average AUPR score and 1% higher than the average AUC (see Table 3) .In Table 3, the top result in every row is italicized. For each data set, our model is in AUPR 0.88 (0.094), 0.86 (0.031),The NR, GPCR, IC, and E data sets are 96 (0.013) and 0.97 (0.005), respectively, and the value in parentheses is 10 times the standard deviation of AUPR in CV. For the GPCR, IC, NR,and E datasets, proposed model outperforms the second best method (namely - TriModel). Our model is also superior to other methods in the following aspects except for TriModel, the AUC conditions of each data set have the same performance for the NR, IC, and E data sets. The best DTI prediction performance is obtained using the IC and E data sets; this can be explained by the fact that these data sets contain a wider set of positive interaction data sets, and the model can use these data to improve the predictions feature. In addition, we use a single AUPR value from an experiment with a tenfold CV to calculate the statistical significance of the performance improvement of our method compared to the suboptimal method.

VI. CONCLUSION

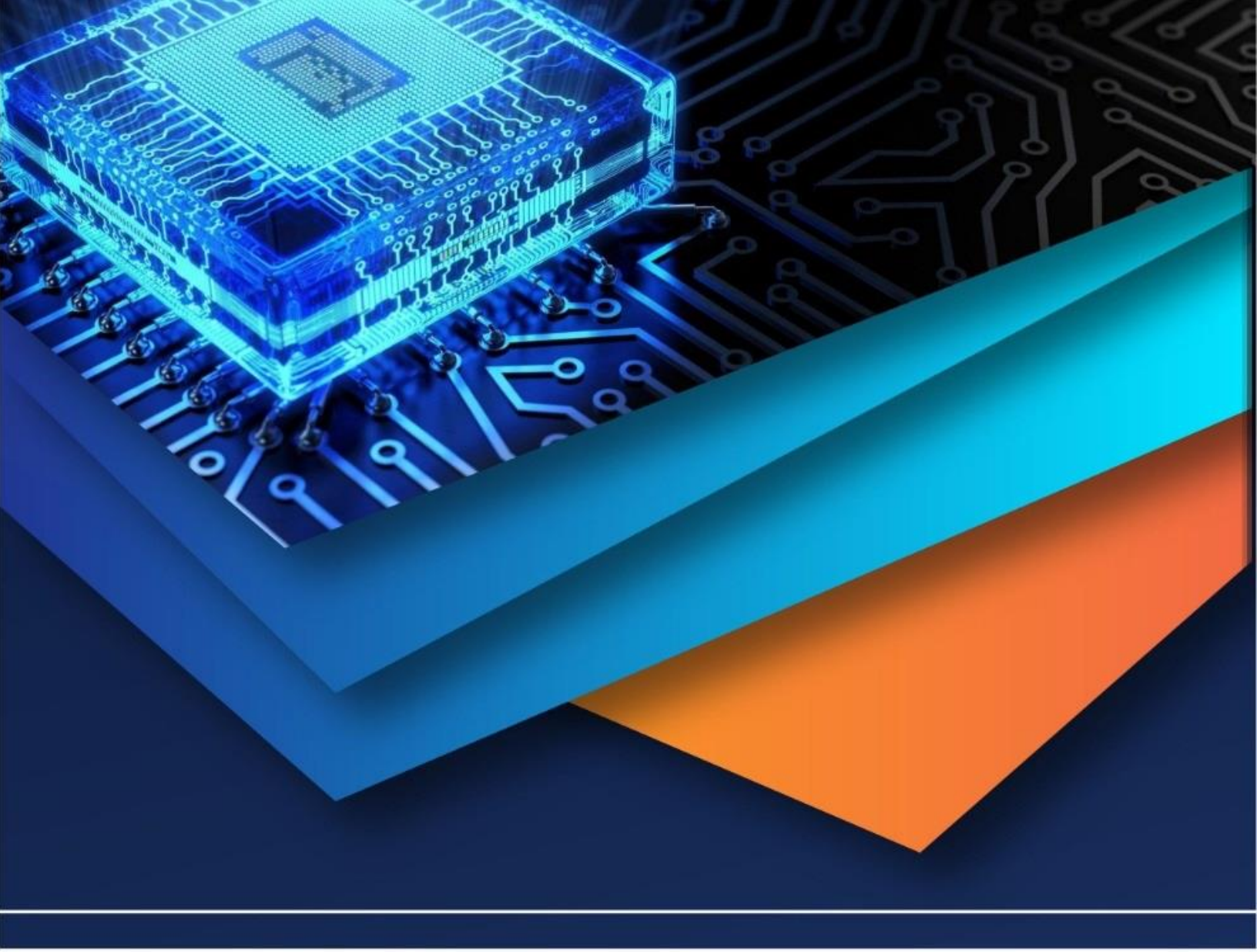
Our work proposes a new computational method that uses deep learning methods to predict drug-target interactions. We integrate various machine learning technologies, graph embedding, mining and similarity-based technologies. That is, (1) use graph embedding in the node2vec drawing function feature to take advantage of network topology and structural features, (2) use graph mining to extract attributes from path estimation, (3) use similarity-based methods to select and integrate information from different sources. Multiple similarities, and finally (4) ML for classification. The novelty of our method is to create path estimation features based on two graphs created with the same DTI but with different types of similarity matrices, where for example, Plot G1 is used to combine the similarity between the drug and the target, providing additional information about the chemical structure and side effects of the drug, as well as the genetic ontology and amino acid sequence of the target protein , And Plot G2 shows the cosine similarity between the active ingredient and the active ingredient and the cosine similarity between the target and the target used to generate topological information. Compared to the seven current methods that use different indicators and predict new DTIs, their performance has been verified using published literature and various online databases. For further improvement, we recommend using different embedding methods, combining more affinity indicators of more fonts, and creating more graphic features for Unrestricted drug and target (unknown interaction), In the future, to further our study we plan to expand the capabilities of our model to create a set of more reliable negative DTIs. In addition, we intend to use our method to predict the DTI of a new drug or new target. Some possible extensions of our work include application to other graphics (such as the web), which are formulated as link prediction problems. Popular examples of linkage prediction in bioinformatics include the prediction of drug interactions, the prediction of drug-disease interactions, and the prediction of gene to disease association. Another extension might be a modification that aims to treat DTI as a regression problem that predicts the binding affinity between a drug and its target protein.

REFERENCES

[1] Pengyong Han, Zhengwei Li, Xiao Li, Yusen Zhang, Zhu-Hong You, Haiquan Yu, Ru Nie & Xing Chen. In silico prediction of DTI network based on drug chemical structure and protein sequence. Scientific Report, 7: 11174, DOI:10.1038/s41598-017-10724-0

[2] María Jesús García-Godoy, Esteban López-Camacho, José García-Nieto, Javier Del Ser, Antonio J. Nebro, José F. Aldana-Montes. Bio-inspired optimization for the molecular docking problem: State of the art, recent results and perspectives. DOI: doi.org/10.1016/j.asoc.2019.03.044

[3] Lucia Perez-Gandarillas, Crina Grosan, Hossam Zawbaa, Abderrahim Michrafy, Serena Schiano, Chuan-Yu Wu. Computational intelligence modelling pharmaceutical tabletting processes using bioinspired optimization algorithms. HAL Id: hal-01944490 https://hal-mines-albi.archives-ouvertes.fr/hal-01944490
[4] Sese J, Tomii K, Tsubaki M; Compound–protein interaction prediction with end-to-end learning of neural networks for graph and sequences; Bio-informatics 35(2):309–318
[5] Rayhan Fet, iDTI-ESBoost: Identification of drug target interaction using evolutionary and structural features with boosting. Sci Rep.7(1):17731
[6] Pathak S, Cai X, Ensemble learning algorithm for drug–target interaction prediction. 2017 IEEE seventh international conference on computational advances in Bio and medical sciences (ICCABS), 2017
[7] Nascimento ACA, Prudencio RBC, Costa IG (2016) A multiple kernel learning algorithm for drug–target interaction prediction. BMC Bioinform(17:46)
[8] Wen Zhang, Ding Zhang, Weiran Lin, Siman Wang, Jingwen Shi and Yanqing Niu. Recent Advances in the Machine Learning-based Drug-target Interaction Prediction.
[9] Ba-Alawi W et al, DASPfind: Efficient method to predict drug–target interactions. J Cheminform 8:15
[10] Yamanishi Y, Prediction of DTI networks from the integration of genomic and chemical spaces, Bio-informatics 24(13): i232-i240.
[11] Luo Yet al, A network integration approach for drug–target interaction prediction and computational drug repositioning from heterogeneous information. Nat Commun. 8(1):573
[12] Nascimento ACA, Prudenciu RBC, Costa IG; A multiple kernel learning method for drug–target interaction prediction. BMC Bio-inform17:46
[13] Smith SB, Identification of common biological pathways and drug targets multiple respiratory viruses based on human host gene expression analysis. PLoS ONE 7(3):e33174
[14] Tenenbaum D KEGGREST: client-side REST access to KEGG. R Package Version 1.24.0. Fred Hutchinson Cancer Research Center
[15] Palme J, Hochreiter S, Bodenhofer U KeBABS: an R package for kernel-based analysis of biological sequences, Bioinformatics 31(15):2574–2576
[16] Hao Ding, IchigakuTakigawa, Hiroshi Mamitsuka and Shanfeng Zhu, Similarity-based machine learning methods for predicting DTIs: a brief review, BRIEFINGS IN BIOINFORMATICS, volume 15:Number- 5: 734 -747
[17] Cai H, Zheng VW, Chang KC, A comprehensive survey of graph embedding- problems, techniques, and applications. IEEE Trans Data Eng 30(9):1616–1637
[18] Ashoor H, Olayan RS, Bajic VB, efficient computational method to predict DTIs using graph mining and ml approaches. Bio-informatics 34(7):1164–1173
[19] Xuan P et al, Gradient boosting decision tree-based method for predicting genes and drug interactions. Front Genet. 10:459
[20] Rawan S. Olayan, Maha A. Thafar, Somayah Albaradei, Xin Gao, Vladimir B. Bajic, Magbubah Essack, Takashi Gojobori and Haitham Ashoor; dtigems: drugtarget interaction prediction using graph embedding, graph mining, and similarity-based techniques, Thafar et al. J Cheminform 12:44 DOI: https://doi.org/10.1186/s13321-020-00447-2
[21] Ruolan Chenn, Xiangong Liu, Shuting Jin, Jiawe Lin and Juan Liu, Machine Learning for Drug-Target Interaction Prediction, Molecules 2018, 23, 2208; doi:10.3390/molecules23092208
[22] Naveed H, An integrated structure and system based framework to identify new targets of metabolites and drugs. Bioinformatics: vol - 31(24):3922–3929
[23] Ezzat A, Computational prediction of drug–target interactions using chemogenomic approaches, Brief Bio-inform 20(4):1337–1357
[24] Sachdev K, Gupta MK, A comprehensive review of feature based methods for drug target interaction prediction. J Biomed Inform 93:103159



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY