Designing of Information Model for Prediction of Drug-drug Interactions based on Calculation of Target and Therapeutic Similarity

Olha Marushchak[a], Rostyslav Kosarevych[a]

[a] *Lviv Polytechnic National University, Stepan Bandera street 12, Lviv, 79016, Ukraine*

Abstract

Understanding and predicting the drug–drug interactions is an important in both drug development and clinical application, especially for co-administered medications. We propose a new information model for drug–drug interaction analysis, based on the common biological targets and therapeutic similarity. Based on the data from DrugBank, our model calculates target similarity and therapeutic similarity features. To predict the possible drug-drug interactions it uses a semi-supervised approach, defined in two steps: adding the missing labels using the clustering algorithm K-Means, and then, executing a classification with a supervised learning model Support Vector Machine. Proposed model is tested for the known data set and had shown the high classification rate, with the AUC=98.5+-0.05.

Keywords 1

Machine learning, predictive model, drug-drug interactions, drug-related data, data analysis, target similarity, therapeutic similarity

1. Introduction

Identification and prediction of drug-drug interactions (DDI) is a widespread topic of the research in the healthcare, and studying such aspect is a big part of the drug development process [1]. Drug-drug interactions occur when two or more drugs react with each other and are vital for the patient safety and success of treatment modalities, they can lead either to the loss of efficacy an adverse drug reaction, or cause the increasing of the therapeutic effect [1, 2]. DDIs can be categorized into three types: pharmaceutical, pharmacokinetic and pharmacodynamic [3].

A number of computational methods have been employed for the prediction of DDIs based on drugs structures and/or functions: physiologically based pharmacokinetic modeling, molecular structural similarity analysis, ontology and annotation-based analysis, network modeling, QSAR modeling [4].

We can divide the machine learning-based methods used for the prediction of DDIs according to the approach used: unsupervised, supervised, and semi-supervised machine learning-based algorithms [21].

In one study [4] it was proposed to use an unsupervised machine learning model for predicting DDIs using the structural similarities of drugs from the Pharmacokinetic and Pharmacodynamic networks and investigated the factors influencing DDIs for further improvement of the predictions. In other study [5], the drug-target pairs were predicted, resulting in a network with strong local clustering of similar types according to Anatomical Therapeutic Chemical (ATC) classification. In other studies, it was used the genomic data and the drug structural characteristics, or the physical and chemical features of drug molecules to create different hypothesis on the possible DDIs and proceed the unsupervised machine learning approaches [8, 10, 11].

IDDM'2020: 3rd International Conference on Informatics & Data-Driven Medicine, November 19–21, 2020, Växjö, Sweden
EMAIL: olha.marushchak.w@gmail.com (O. Marushchak); kosar2311@gmail.com (R. Kosarevych)
ORCID: 0000-0001-5620-1299 (O. Marushchak); 0000-0001-9108-0365 (R. Kosarevych)
© 2020 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-2753/short11.pdf

The aforementioned studies proceeded a vast amount of data [4,8,19]. The research objectives mostly included investigating the underling mechanisms of possible drug-target and drug-drug interactions [11, 19, 20]. However, it was noticed that the known DDIs were not taken into consideration for the unsupervised learning models [4, 5, 11]. We assume that the known DDIs are a valuable piece of information, as their characteristics can serve as a benchmark for not yet discovered combinations.

Several studies focused on predicting the DDI though protein-protein-interaction networks with chemical features [3, 6, 7], implementing supervised learning models to the labeled data. The similarity-based approach has been used to predict the possible outcomes of combining drug pairs [10, 14, 15].

Only few of the studies proceeded the 'in vitro' experiments to evaluate their models [9, 11, 12], other evaluated the performance of their methods by comparison of the predicted DDI and the reported ones in the literature [4, 14, 16].

We have observed that the supervised learning models analyzed significantly smaller amount of possible DDIs [15, 16]. It was caused by the relatively small number of the known DDIs in the literature, and therefore, picking the corresponding amount of the drug pairs without indication of possible interaction in the literature. Besides, we assume that there should be more complex procedure of data labeling, because it improves the performance of the next predictions [12].

So, predicting DDIs is a complex problem [1, 3] that requires addressing it from the medical perspective – in a form of creating a hypothesis and picking suitable drug-related characteristics; and from the computer science perspective – by choosing the appropriate computational methods and predictive models.

In this study, we propose a new information model for drug–drug interaction analysis, based on the common biological targets and therapeutic similarity. The information model is able to proceed the data extraction from the source, execute the calculation of the features, execute the data labeling process and make predictions of the possible DDIs.

Regarding that many researchers obtained data for their investigations of DDI from a database DrugBank [17], we used DrugBank as the data source for our work as well. Also we used the calculation methods of target and therapeutic similarities features proposed by Cheng et al. [l5, 16] – such approach, combined with the supervised learning algorithm, Support Vector Machine has shown a significant accuracy in predicting DDI on the sample.

We examined the hypothesis of predicting the based on the common biological targets and therapeutic use, instead of including chemical and physical descriptors of the drugs.

We addressed the following problem: the researchers added labels meaning the absence of DDI when the drug pair didn't have the DDI indication in the data source, however there might be the unreported, and used only 3% of the original input data. In this study we want to improve the data labeling process, and that would enable us to use the whole dataset as well for the next predictions. So, we followed a semi-supervised approach, which consists of first, clustering algorithm for obtaining the missing data labels, and then, classification to predict the possible drug-drug interactions.

2. Methods and Materials

2.1. Obtaining Input Data

We used a database DrugBank as a source of data. Drug bank is freely accessible, online database containing information on drugs and drug targets [17]. It contains both bioinformatics, cheminformatics details about drugs such as resource, chemical, pharmacological and pharmaceutical data with comprehensive drug target (sequence, structure, and pathway) information.

Obtaining data for the study was performed by parsing the xml document. We extracted the following drug information: drug name, DrugBank ID of the drug, targets, ATC code (anatomical-therapeutic-chemical classification), known drug-drug interactions. From the obtained set of drugs and characteristics, we removed drugs that did not contain ATC code (experimental drugs, homeopathic and herbal traditional medicinal products) as well as drugs of antibodies and inorganic salts.

2.2. Designing of the Information Model

The information model was designed to proceed all the steps required for the drug-related data analysis and prediction of the possible drug-drug interactions.

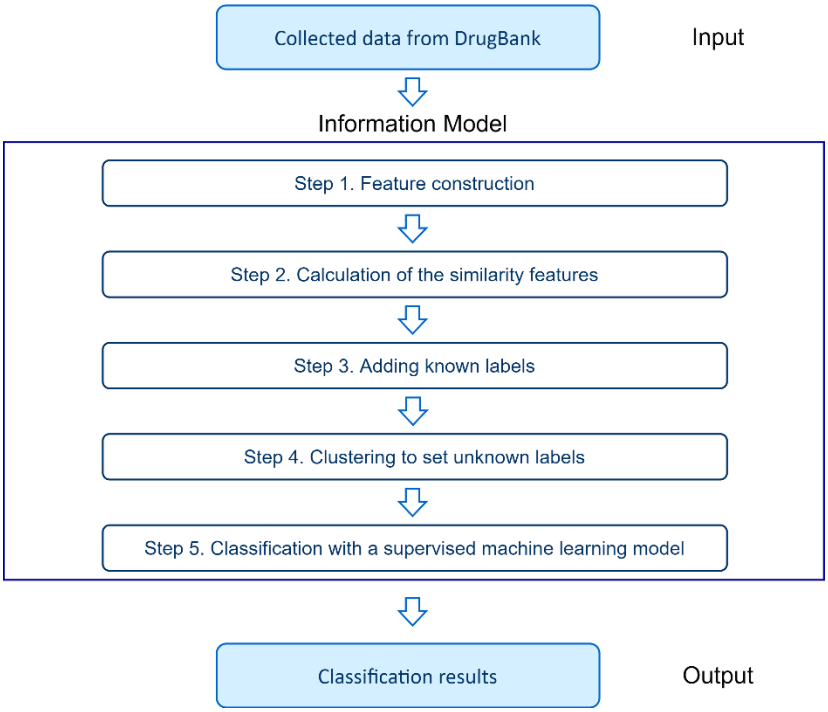

Figure 1: Schema of the process held within the information model

All possible unique drug combinations were composed, and the calculation of the target similarity and therapeutic similarity features was performed.

For the target similarity (SB) feature we used an approach proposed by Cheng et al. [16]. We summarized all the unique biological targets identified for the drugs, added them to the general sequence, and created binary vectors for each drug. If the drug affects a biological target - then a certain element of the vector contains a value of 1, if the drug has no effect on the target - the value is 0. After that, for each combination of drugs, we constructed the target similarity by calculating the Tanimoto coefficient for binary vectors of drugs:

$$SB(a,b) = \frac{N_{ab}}{N_a + N_b - N_{ab}}, \quad (1)$$

where N_{ab}represents the quantity of the biological targets, which are common for both drugs of the combination;
$N_a$represents quantity of the biological targets, which the drug a affects;
$N_b$represents quantity of the biological targets, which the drug b affects.

For the therapeutic similarity feature (ST) we used the method proposed by Cheng et al. [15]. We created five sets with unique ATC codes representing each of the five ATC classification levels for each drug pair. Next, for each drug pair for each ATC classification level the therapeutic similarity feature was calculated:

$$ST_k(a,b) = \frac{ATC_k(a) \cap ATC_k(b)}{ATC_k(a) \cup ATC_k(b)}, \quad (2)$$

where k represents an ATC classification level (from 1 to 5);
$ATC_k(a)$ represents ATC codes of the k-level for the drug a;
$ATC_k(b)$ represents ATC codes of the k-level for the drug b.

After that, the general therapeutic similarity was calculated considering all five ATC classification levels:

$$ST(a,b) = \frac{\sum_{k=1}^{n} ST_k(a,b)}{n}, \quad (3)$$

where n represents the overall number of ATC classification levels (=5);
$ST_k(a, b)$represents the previously calculated therapeutic similarity for each ATC classification level.

The drug pairs that were indicated in the DrugBank as known, received the label 1. For the remaining drug combinations there is not enough information in DrugBank to assert or deny the drug-drug interaction, so no labels were added.

In order to predict the labels for the drug pairs that contained no label, the clustering method K-Means was applied. We have chosen K-Means method because it has been used by the researchers in the healthcare fields as the first step of the semi-supervised machine learning approach to define the missing data labels [10, 18].

In K-Means method the centroids are randomly initialized from the dataset. Then, from each centroid the Euclidean distance is calculated to each data point, and depending upon the minimum distance between the centroids and data points, that data point is assigned to that centroid. This is repeated until there is no change of the centroids. In this way, the clusters are formed.

The accuracy of the clustering was calculated according to the percentage of how much of the clustered labels 1 match the original labels 1 for the drug pairs.

For the predicting of the possible drug-drug interactions, we used the supervised machine learning model Support Vector Machine (SVM), namely Linear Support Vector Classification. We made our choice based on the literature review: in drug-related research this method is used to solve classification problems. In the studies we investigated, such method has shown significantly good performance [14, 16].

The AUC value was calculated, and the confusion matrix was composed to evaluate the performance of the model.

The research was performed using the programming language Python3. The xml parsing was proceeded using the library by using library xml.etree.ElementTree in Python3. The data analysis was performed using libraries numpy, pandas, sklearn. Data visualization was executed using libraries matplotlib, seaborn.

We used open-source software which is freely available and contributed by the global community of developers.

3. Results

From DrugBank we obtained information about 721 drugs, which has been used as an input for the information model.

For the feature construction, 266085 unique drug-drug combinations were created. The target and therapeutic similarities were calculated and assigned to the corresponding drug pairs. 6946 drug pairs were actually indicated in DrugBank as having the drug-drug interaction, so they received the label 1.

Whole dataset was used as an input for the clustering algorithm, with k=2 clusters. The accuracy calculated with our method is 54%.

After that, for all known drug combinations that contained the label 1 before clustering, we left the original labels, and for the drug combinations with the missing ones, we assigned the labels obtained as a result of clustering.

We investigated the distribution of each feature according to their labels.

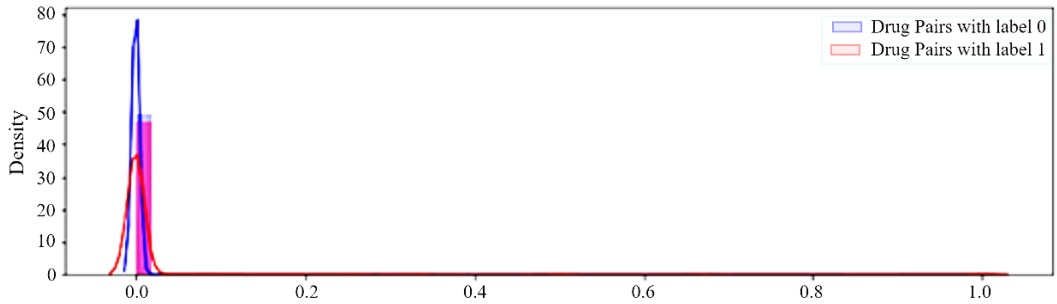

Figure 2: Distribution of the feature Target Similarity

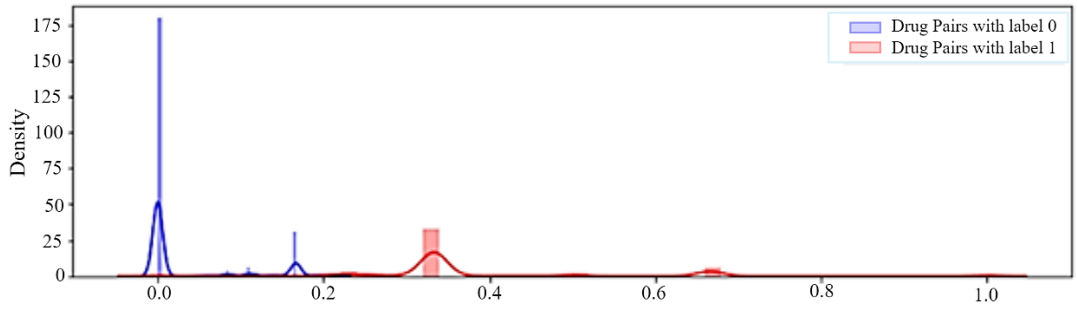

Figure 3: Distribution of the feature Therapeutic Similarity

The distribution is binomial, the values are contained in range [0, 1]. It is noted that the drug combinations with label 1 have two density peaks in the area 0.3 and 0.65; and the drug combinations with label 0 have two density peaks in the area 0 and 0.175.

To execute the classification, the whole dataset was splitted into the train set (70%) and test set (30%). Such splitting ratio (70/30) has been widely used by the scientific community for data analysis. In our research we didn't notice the significant difference of performance with various splitting, but with the 70/30 ration the AUC value was the highest – 98.53 (Comparing to AUC=98.39 for 90/10, AUC=98.41 for 80/20, AUC=98.47 for 60/40).

We applied the Linear Support Vector Classification method of the Support Vector Machine algorithm with the linear kernel.

Based on the prediction, we received the Area Under the Curve (AUC) value = 98.5+-0.05 illustrate the absolute values of prediction of the training set, we composed the following confusion matrix:

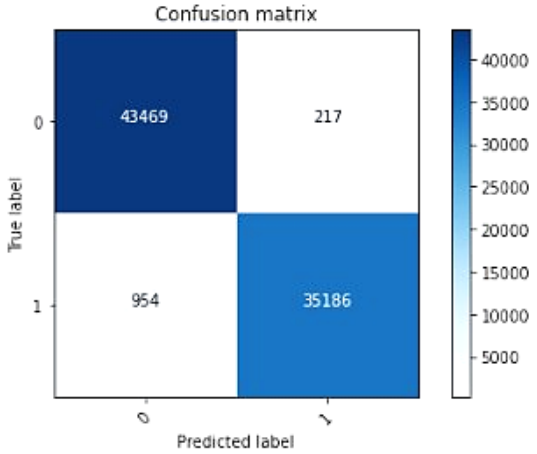

Figure 4: Confusion matrix to evaluate the predictions of information model

The count of predicted drug-drug interactions is 35 186. The number of correctly predicted drug pairs that do not have drug-drug interactions is 43 469. There are 217 drug pairs with label 1 that were predicted as such not having the drug-drug interaction. 954 drug pairs were wrongly predicted as having the drug-drug interaction, although they were labeled as 0.

4. Conclusion

The information system for the drug-related data analysis and the prediction of the possible drug-drug combinations based on their calculated target and therapeutic similarities was created. It uses a semi-supervised learning approach, in order to firstly, define the missing labels using the clustering algorithm, and then, execute a classification using a supervised learning model.

Our examined hypothesis to use data about biological targets and therapeutic use has received reinforcement in the form of high predictive performance on the dataset from DrugBank, verified with the test set.

By executing the data labeling process, we were able to use for the further predictions all amount of drug combinations, including the 97.39% that didn't have the labels at the beginning.

In the similar studies that use same dataset, included biological targets or therapeutic use into their examined hypothesis, the results were AUC=0.968 [14] and AUC=0.912 [15]. So, our implementation of the proposed information system has shown accuracy of classification about 98.5+-0.05 (AUC) for the DrugBank dataset and it outperforms other similar systems.

This information system can be enhanced with the functionality to calculate more features such as enzyme similarity and transporter similarity.

5. Acknowledgements

We thank Olga Boretska (Danylo Halytskyi Lviv National Medical University) for providing insight and expertise that greatly assisted the research.

6. References

[1] Waters, Nigel J. "Evaluation of drug–drug interactions for oncology therapies: in vitro–in vivo extrapolation model-based risk assessment." British journal of clinical pharmacology 79.6 (2015): 946-958.

[2] U.S. Food and Drug Administration: Drug Interaction: what you should know. URL: https://www.fda.gov/drugs/resources-you-drugs/drug-interactions-what-you-should-know

[3] Huang, Jialiang, et al. "Systematic prediction of pharmacodynamic drug-drug interactions through protein-protein-interaction network." PLoS Comput Biol 9.3 (2013): e1002998.J. Cohen (Ed.), Special issue: Digital Libraries, volume 39, 1996.

[4] Takeda, Takako, et al. "Predicting drug–drug interactions through drug structural similarities and interaction networks incorporating pharmacokinetics and pharmacodynamics knowledge." Journal of cheminformatics 9.1 (2017): 1-9.

[5] Yıldırım, M., Goh, K., Cusick, M. et al. Drug—target network. Nat Biotechnol 25, 1119–1126 (2007). https://doi.org/10.1038/nbt1338.

[6] Lei Huang, Fuhai Li, Jianting Sheng, Xiaofeng Xia, Jinwen Ma, Ming Zhan, Stephen T.C. Wong; DrugComboRanker: drug combination discovery based on target network analysis, Bioinformatics, Volume 30, Issue 12 (2015): i228–i236

[7] Zhao, Xing-Ming, et al. «Prediction of drug combinations by integrating molecular and pharmacological data» PLoS computational biology 7.12 (2011): e1002323.

[8] Chandrasekaran, Sriram, et al. "Chemogenomics and orthology-based design of antibiotic combination therapies." Molecular systems biology 12.5 (2016): 872.

[9] Li, Xiangyi, et al. "Prediction of synergistic anti-cancer drug combinations based on drug target network and drug induced gene expression profiles." Artificial intelligence in medicine 83 (2017): 35-43.

[10] Ferdousi, Reza, Reza Safdari, and Yadollah Omidi. "Computational prediction of drug-drug interactions based on drugs functional similarities." Journal of biomedical informatics 70 (2017): 54-64

[11] Aghakhani, Sara, Ala Qabaja, and Reda Alhajj. "Integration of k-means clustering algorithm with network analysis for drug-target interactions network prediction." International Journal of Data Mining and Bioinformatics 20.3 (2018): 185-212.

[12] Chen, Xing, et al. "NLLSS: predicting synergistic drug combinations based on semi-supervised learning." PLoS computational biology 12.7 (2016): e1004975.

[13] Peng Li, Chao Huang, Yingxue Fu, Jinan Wang, Ziyin Wu, Jinlong Ru, Chunli Zheng, Zihu Guo, Xuetong Chen, Wei Zhou, Wenjuan Zhang, Yan Li, Jianxin Chen, Aiping Lu, Yonghua Wang; Large-scale exploration and analysis of drug combinations, Bioinformatics, Volume 31, Issue 12, 15 June 2015, Pages 2007–2016

[14] Song, Dalong, et al. "Similarity-based machine learning support vector machine predictor of drug-drug interactions with improved accuracies." Journal of clinical pharmacy and therapeutics 44.2 (2019): 268-275.
[15] Cheng, F., Li, W. et al. Prediction of polypharmacological profiles of drugs by the integration of chemical, side effect, and therapeutic space. Journal of chemical information and modeling, 53(4), (2013): 753-762.
[16] Cheng, F., & Zhao, Z. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. Journal of the American Medical Informatics Association, 21(e2), (2014): e278-e286.
[17] Wishart, David S., et al. "DrugBank: a knowledgebase for drugs, drug actions and drug targets." Nucleic acids research 36.suppl_1 (2008): D901-D906.
[18] Singh, Reetu, and E. Rajesh. "Prediction of Heart Disease by Clustering and Classification Techniques." International Journal of Computer Sciences and Engineering 7 (2019): 861-866.
[19] Li, Xiangyi, et al. «Biomolecular network-based synergistic drug combination discovery» *BioMed research international*2016 (2016).
[20] Wu, Zikai, Xing-Ming Zhao, and Luonan Chen. «A systems biology approach to identify effective cocktail drugs» *BMC systems biology*. Vol. 4. No. 2. BioMed Central, 2010.
[21] Li, Xiangyi, et al. "Biomolecular network-based synergistic drug combination discovery." *BioMed research international* 2016 (2016).