IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 19, 2019, accepted March 18, 2019, date of current version April 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2907522

Drug-Disease Association Prediction Based on Neighborhood Information Aggregation in Neural Networks

YINGDONG WANG[1], (Member, IEEE), GAOSHAN DENG[2], NIANYIN ZENG[3], XIAO SONG[4], AND YUANYING ZHUANG[5]

[1]Software School, Xiamen University, Xiamen 361005, China
[2]Computer Science Department, University of Southern California, Los Angeles, CA 90089, USA
[3]Department of Instrumental and Electrical Engineering, Xiamen University, Fujian 361005, China
[4]School of Computer and Information Technology, Nanyang Normal University, Nanyang 473000, China
[5]School of Mathematics and Statistics, Nanyang Institute of Technology, Nanyang 473000, China

Corresponding authors: Nianyin Zeng (zny@xmu.edu.cn) and Yuanying Zhuang (yuanying566@foxmail.com)

This work was supported by the National Natural Science Foundation of China (the Joint Funds of Henan Province) under Grant U1504605. This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

ABSTRACT Computational drug repositioning plays a vital role in the prediction of drug function. Many new functions discovered have been confirmed. In comparison with traditional drug repositioning, computational drug repositioning shortens the time and reduces labor. Thus, it has received wide attention in recent years. However, prediction remains a considerable challenge. In this paper, a method called HNRD is introduced to predict the link between drugs and diseases. It is based on neighborhood information aggregation in neural networks which combines the similarity of diseases and drugs, the associations between the drugs and diseases. Compared with the state-of-the-art method before, our method has achieved better results, with the best AUC of 0.97 in one of the golden datasets. To better evaluate our approach, we also performed data analysis based on one-to-one association's prediction and robust analysis by testing on different datasets. All the results prove the excellent performance of prediction. Source codes of this paper are available on https://github.com/heibaipei/HNRD.

INDEX TERMS Drug reposition, deep learning, matrix decomposition, heterogeneous network, end to end.

I. INTRODUCTION

Drug research and development is a complex, lengthy and expensive process. It often takes 10-15 years of research and 0.8-15 billion dollars to make a drug from abstract concept to market-ready product [1]. Annually, 90% of drugs fail to get access to FDA evaluations, thereby preventing their use in actual therapy [2]–[5]. Accordingly, Drug Repositioning (DR) based on computing method appears. The repositioning method bypasses many pre-approval tests that are critical to newly developed therapeutic compounds, and it can shorten the drug development cycle to 3-12 years for a repositioned drug [6]. In recent years, DR has received increased interest from governments, nongovernmental agencies and academic researchers.

In general, DR seeks to find new uses for existing drugs, with established and demonstrated human safety. In technical terminology, DR is the process by which new indications are found for approved drugs [7]. Recently, the usage of computational DR in drug discovery has become a popular practice, and an increasing number of machine learning [8]–[10], network analysis [11]–[13], text mining and semantic inference methods [14] have been proposed [15]–[20].

PREDICT [21] calculates the link between potential drugs and diseases, mainly by integrating the similarities between various drugs, diseases, and using these features to obtain new potential features through a logical classifier. DRRS [22] merges three matrices, including the drug similarity matrix, the disease similarity matrix, and the drug and disease association matrix, into one large matrix. Then it finds the lowest level of the big matrix that reconstructs the large matrix. NeoDTI [23] predicts new drugs and drug targets by integrating various information in heterogeneous networks and conducting end-to-end learning through a nonlinear model. TL-HGBI [24] has proposed a computational framework, to infer novel treatments for diseases based on a heterogeneous network integrating similarity and association data about diseases, drugs and drug targets. DrugNet [25] has

The associate editor coordinating the review of this manuscript and approving it for publication was Ying Song.

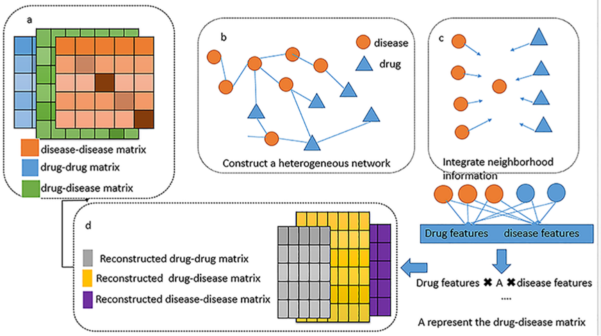
a
disease-disease matrix
drug-drug matrix
drug-disease matrix
b
disease
drug
Construct a heterogeneous network
c
Integrate neighborhood information
Drug features
disease features
Drug features ✖ A ✖disease features
....
A represent the drug-disease matrix
d
Reconstructed drug-drug matrix
Reconstructed drug-disease matrix
Reconstructed disease-disease matrix

FIGURE 1. **HNDR flowchart: (a) Input data include a reconstructive drug-drug similarity matrix, a disease semantic similarity matrix, and an experimentally verified drug-disease association matrix. The similarity matrices are symmetric. (b) Construct a heterogeneous network. Each node of the network is either drug or disease and is initialized with a low dimensional vector representation. (c) Perform neighborhood information integration, which updates nodes representation. (d) Reconstruct three matrices with learned node representation. These three new matrices serve the next input of feature extraction procedure. The procedure is designed to minimize the difference between the initial matrices and reconstructed matrices, the reconstructed drug-disease matrix is used to predict potential associations between drugs and diseases.**

developed a network-based prioritization method to predict new therapeutic indications for drugs and novel treatments for diseases. This method identifies novel drug-disease associations by propagating information in a heterogeneous network which is constructed by using all the information about diseases. Reference [26] integrates miRNA similarity and disease similarity based on the functional similarity of miRNA, disease semantic similarity and Gaussian interaction profile kernel similarity, and predicts the association between miRNA and disease through inductive matrix completion. Reference [27] has proposed MBiRW utilizing some comprehensive similarity measures, and Bi-Random walk (BiRW) algorithm to identify potential novel indications for a given drug. By integrating information about drug or disease features with known drug-disease associations, the comprehensive similarity measures are initially developed to calculate the similarity between drugs and diseases, which has demonstrated certain success in computational DR, and other drug or disease association [28]–[35]. Although some of these methods are predictions of the potential relationship between drugs and drug targets [36], [37], all methods prove that multiple relationships integrated into a graph could improve the effect of prediction novel link.

Inspired by currently popular neural-network-based approaches, we introduce a neural-network-based method of neural-network-based integration of neighborhood information in a Heterogeneous Network for drug-Disease association prediction (HNRD) in this paper to predict novel associations between drug and disease. In the HNRD, a heterogeneous network is first generated from the dataset with each node, either drug or disease, by integrating neighborhood information, which is achieved through the nonlinear feature learning. Then, HNRD enforces the embedding node representations of drugs and diseases to match the observed matrices. HNRD is a global approach that can rank candidate drug-disease pairs for all diseases simultaneously. In tenfold cross-validation experiments, our method achieves an area under the receiver operating characteristic curve (AUC) of 0.97 in one of the golden datasets, which is approximately higher than that of the state-of-the-art method. Additionally, we perform leave-one-out cross-validation (LOOCV) experiments based on the new drug prediction where only one association exits drug and disease; our method achieves a perfect result, which is approximately 2% higher than the state-of-the-art method we know. Finally, to further prove the validation we use different dataset with our algorithm.

The main contribution of this paper involves the following: (1) our proposed method preformed a deep network to extract drug features from the drug-drug matrix and the drug-disease matrix, to extract disease features from the disease matrix and disease-drug matrix, finally, based on the principle of matrix decomposition, the two recessive features are decomposed as the matrix to recover the matrix. It is the first time to apply in the drug-disease reposition using the end-to-end method to recover method. (2) Considering the characters of the network model, large data is needed to train the model. A large amount of data corresponds to a good AUC. Among the datasets, the DNdataset has the biggest matrix, which has the highest AUC of 0.972. Its precision rate can reach 0.802, which is much higher than that of the state-of-the-art method with an AUC of 0.935 and maximum precision is 0.348. It would not be influenced by the sparse as long as the amount of data is abundant.

II. MATERIALS AND METHODS

In this study, we propose a novel DR HNRD approach to infer potential drug indications. First, we provide a brief description of our datasets. Then, HNRD is utilized to train the prediction model to predict the missing association in the test dataset.

A. DATASETS

The gold standard dataset include three apartments. For the drug-drug similarity matrix, the chemical structure of all drugs are download from DrugBank in the Canonical Simplified Molecular-Input Line-Entry System (SMILES) format [38], and then a two-category is calculated according to the Chemical Development Kit [39]. Finally, based on the two fingerprints the similarity is calculated, with a range of [0, 1].

For the disease-disease similarity matrix, a phenotype-based disease-disease similarity dataset is downloaded from MimMiner [40], which was constructed by calculating similarities based on the numbers of occurrences of Medical Subject Headings vocabulary (MeSH) terms in the medical descriptions of each pair of diseases from the OMIM database [41]. According to the MimMiner database description, the similarities have already been normalized to the range [0, 1].

For the drug-disease matrix, initial disease drug interactions were obtained from [21], where disease and drug interactions are assembled for the diseases listed in the OMIM

TABLE 1. Statistics of the gold standard dataset used in this study. Sparsity is defined as the ratio of the number of known interactions to the number of all possible interactions.

Dataset	Drugs	Disease	Interaction	$Sparsity^{a}$
Fdatasets	593	313	1933	1.041^{-2}

database and their associated drugs but are limited to the ones registered in the DrugBank database [42]. The corresponding value in the matrix W_{dr} was set to 1 if an interaction exists and 0 otherwise.

B. SCHEMATIC OVERVIEW OF HNRD

a) Construct a heterogeneous network based on three standard matrices. The three matrices mainly include the drug similarity link matrix, the disease similarity adjacency matrix, and the correlation matrix of drugs and diseases. The similarity matrix is symmetrical, whereas the drug-disease correlation matrix is asymmetric and binary. Regularize the correlation matrix for each pair. b) Integrate neighborhood information for drugs and diseases, and embed low-dimensional space, each with a low-dimensional representation. c) Reconstruct the drug-disease matrices with the captured feature vectors. This step is intended to minimize the different between the reconstructed matrices and the initial matrices. It also can be considered as an embedding process to maximize the extraction of information about the three matrices. e) Finally, predict the drug-disease sequence by reconstructing the matrix. The whole task can be considered as a filling of the matrix, mainly to fill the data that is not in the part [43].

C. HETEROGENEOUS NETWORK

Let $S_{rr} \epsilon R^{m\times m}$ be the drug expression profile similarity matrix and $S_{dd} \epsilon R^{n\times n}$ be the similarity matrix between diseases. Let $A \epsilon R^{m\times n}$ be the drug-disease association matrix, where for each number a_{ij} in A, $a_{ij} = 1$ if $drug_{(i)}$ is connected to $disease_{(j)}$, otherwise, $a_{ij} = 0$. Elements of each matrix are non-negative. For each matrix, we conduct normalization before further processing. Let $S_{rr'}$, $S_{dd'}$, A' be the normalized matrix of drug expression profile similarity matrix, disease semantic similarity matrix and drug-disease associations matrix, respectively such that:

$$M'\{i,j\} = \frac{M\{i,j\}}{\sum_{(k=1)}^{(num(col))} M\{i,k\}} \tag{1}$$

where M stands for matrix and num(col) is the size of the matrix's column dimension. By using the normalized matrices as edges weight, a heterogeneous network is generated which contains two node types {drug, disease} and three edge types {drug-drug, disease-disease, drug-disease}.

D. NODE EMBEDDING

For each node v, drug or disease. Its features should be aggregated from its neighbors, which have a positive weight of connections between node v and its neighbors:

$$R_{e_i}' = concat(R_{ei} \sum_{j=1}^{m} S_d{}'\{i,j\} \times \sigma_{rr}{}^{j} + \sum_{j=1}^{n} A'\{i,j\} \times \sigma_{rd}{}^{j}) \tag{2}$$

$$D_{e_i}' = concat(D_{ei} \sum_{j=1}^{m} S_d{}'\{i,j\} \times \sigma_{dd}{}^{j} + \sum_{j=1}^{n} A'\{i,j\} \times \sigma_{dr}{}^{j}) \tag{3}$$

where $R'_{ei} \epsilon R^{2d}$ and $D'_{ei} \epsilon R^{2d}$ are the embeddings of $drug_i$ and $disease_i$, respectively. The initial representations of nodes ($R_{ei} \epsilon R_d$ or $D_{ei} \epsilon R_d$) are randomly set. Through neighborhood aggregation, we obtain the representation of each node, considering its relation with its neighbor nodes considering its connection and its own nodes features, and we learn the structural and topological information as the feature vectors σ_{xy}^{j} is defined as follows:

$$\sigma_{dr}^{j} = \sigma(\bar{y} e_j W_{xy} \pm b) \tag{4}$$

where W and b are parameters trained in the neural network. $\sigma[\cdot]$ (implemented as $RELU(x) = max(x, 0)$) stands for the activation function in the neural network. In this step, the model further learns node representations into lower dimensional vectors and implement normalization:

$$\sigma_i{}'' = \frac{\sigma(e_j' W_{xy} \pm b_0)}{\|e_j' W_{xy} \pm b_0\|_2}) \tag{5}$$

where e_i'' stands for either R_{ei}'' or D_{ei}''. In this step, a new embedding is learned in a single-layer neural network which non-linearly transforms the representation of the nodes.

E. TRAINING AND EVALUATION

We train the neural network to minimize losses between reconstructed matrices and the initial matrices.

$$\begin{aligned} Loss = {} & \sum (A\{i,j\} - R_{ei}{}'' E_{rd1}^{i} E_{rd2}^{j}{}^{T} De_j'')^{2} \\ & + \sum (Sim_{rr}\{i,j\} - R_{ei}{}'' E_{rr}^{i} E_{rr}^{j}{}^{T} Re_j''^{T})^2 \\ & + \sum (Sim_{dd}\{i,j\} - R_{ei}'' E_{dd}^{i} E_{dd}^{j}{}^{T} De_j''^{T})^2 \end{aligned} \tag{6}$$

Here $E \epsilon R^{d\times k}$ functions as projection matrices, which extract the principle features from node representations. The inner product of the two projected vectors should be reconstructed by the original edge weights as much as possible. For a symmetric matrix reconstruction (drug-drug similarity matrix or disease-disease similarity matrix), the matrix EE^T is used to enforce symmetry of the recovery. A similar reconstruction strategy has also been used [44] to solve prediction problems.

Considering that all operations are differentiable and sub differentiable, parameters can be trained in an end-to-end manner by performing gradient descent. After training, each LDA score could be predicted using the reconstructed drug-disease association matrix. A high score corresponds to a high probability, and we suggest that the following potential association exists:

$$A\{i,j\}_{recovered} = R_{ei}{}'' E_{rd1}^{i} E_{rd2}^{j}{}^{T} De_j'' \tag{7}$$

In this sense, the HNRD prediction task can be considered as a matrix completion problem, which is conventionally solved by matrix factorization with mathematical calculation. By comparison, our method develops a deep learning model to generate feature matrices by explicitly defining the construction process. Through representation learning, HNRD incorporates prior knowledge of network topology, after which the loss minimization procedure is implemented to prevent the network from being arbitrarily factorized. As a result, the method obtains performance improvement in identifying LDA associations.

III. RESULTS

In this section, we systematically evaluate the performance of HNRD by using the datasets. First, the evaluation metrics used in this study are introduced. Then, we compare HNRD with several other methods in terms of prioritizing candidate diseases for a given drug. Next, a case study is conducted to further illustrate the practical usefulness of HNRD. Finally, we perform prediction on the other dataset to verify the robustness of our method.

A. EVALUATION METRICS

To evaluate the effect of HNRD on DR, a ten-fold cross-validation was used. The gold has links to 1933 already known and other unverified links. All known links and the unverified data set are randomly divided into 10, from each of the positive and negative samples as a testset, and the rest as training sets.

When the probability of connection between the drug and disease is re-estimated, the tested links and the candidate links are reordered for each drug. For each specific threshold, four values of true positive (TP), false positive (FP), false negative (FN), and true negative (TN) are calculated [45]–[47]. Predicted value ranks that exceed the threshold are considered correct. TP and TN indicates that the positive and negative samples are correctly predicted, and FN and FP are predicted to be incorrect for the positive and negative samples, respectively. The TPR, FPR, and correct rate are calculated by varying thresholds, resulting in the Receiver Operating Characteristic (ROC) and Precision Rate(PR). For the ROC curve, FPR and TPR are plotted on the x-axis and y-axis, respectively. For PR curve, recall is plotted on the x-axis, and precision is plotted on the y-axis [48]. The area under ROC curve (AUC) value and precision are utilized to evaluate the overall performance of the prediction methods. PR does not represent the preparation rate, but only the existing link probability is ranked first, and the position is ranked later.

$$TPR = \frac{TP}{TP + FN} \tag{8}$$

$$TPR = \frac{FP}{FP + TN} \tag{9}$$

B. COMPARISON WITH OTHER METHODS

To assess the performance of HNRD, we compare it with the other five methods: DRRS [22], MBiRW [27], DrugNet [25], HGBI [24], and KBMF [49]. DRRS constructs a big matrix combine the drug-drug similarity, disease-disease similarity and drug-disease matrix, and finds the minimum rank to reconstruct the drug-disease matrix. MBiRW utilizes the comprehensive similarity measures and Bi-Random Walk algorithm to identify potential novel indications for the given drug. DrugNet is a generic network-based drug repositioning method, which propagates information between networks and can be used to perform both drug-disease and disease-drug. HGBI is introduced based on the guilt-by-association principle and an intuitive interpretation of information flow on the heterogeneous graph. All the parameters used in these methods are determined according to their literature. KBMF is a kernelized Bayesian matrix factorization method, that may work with multiple data side information and can be applied in recommendation systems, the parameter R used as 40 is the same as DRRS.

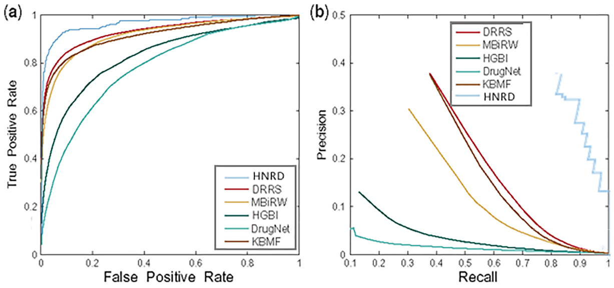

FIGURE 2. (a) Comparison of predicting methods in terms of AUC on the dataset. When the parameter $\alpha = 0.5$, i.e. to wrongly predict unknown entry as positive entry (0 to 1) would cause the same loss as wrongly predict positive entry as negative entry (1 to 0), our method (blue) has an AUC value of 94.2% which is higher than the AUC value (93%) of the state-of-the-art method (red). The other colors indicate the performance of other methods. (b) Comparison of predicting methods in terms of precision and recall, the best value can be 0.562.

The overall performance of all methods is evaluated by applying ten-fold cross-validation. The experiment results in terms of ROC curves and PR curves are depicted in Figure2. Experiment results show that our proposed method outperforms other competitive methods in terms of AUC and precision values. HNRD can achieve an AUC value of 0.942, while the best precision can be 0.572, indicating that it can successfully prioritize 57.2% true drug-disease associations as the ones with the highest rank.

C. PREDICTING INDICATIONS FOR NEW DRUGS

LOOCV was implemented on the known experimentally verified drug-disease associations to evaluate the performance of HNRD. For a given disease d_i, each known drug associated to d_i is left out in turn as the test sample, while the other known experimentally verified drugs associated with d_i are considered as training samples. All the drugs without known associations with d_i make up the d_i-associated candidate samples. In the candidate samples, the test sample is deemed as a positive sample, and the others are negative samples. In each turn, predicting score was recovered by the HNRD method. After all drug-disease entries have been predicted, a special ranking

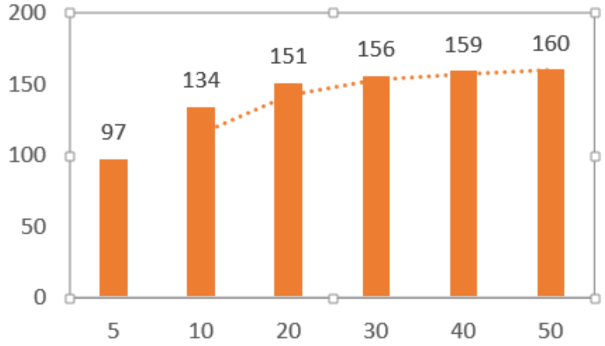

FIGURE 3. Top k associations predicted by HNRD (red) for each disease. In the condition that k =10, 20, 50 or 100, HNRD fetch more corrected association than the state-of-the-art method.

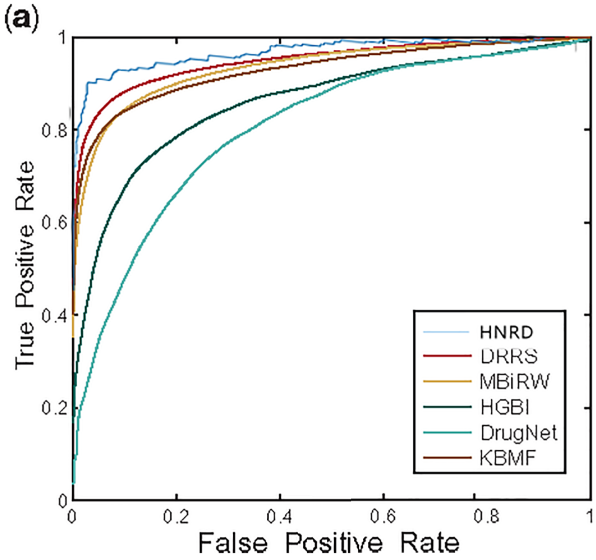

FIGURE 4. Cdataset test results (the blue color) in terms of AUC with the other algorithms. The best result is 0.95.

cutoff was selected as a threshold. Entries with values higher than the threshold are identified as having associations. TPR (sensitivity) measures the proportion of positives that are correctly identified, while FPR (1-specificity) is the percentage of negative samples incorrectly identified.

A total of 171 drugs have only one known disease associations. To make a comparison with the state-of-the-art method, we analyze the performance of all methods for drugs, which have only one known disease association in the dataset. Figure 4 represent the ROC curves. HNRD has achieved superior performance over the other methods. For example, HNRD achieves an AUC value of 0.85, while DRRS, MBiRW, HGBI, DrugNet and KBMF obtain inferior AUC values of 0.842, 0.818, 0.746, 0.759 and 0.806, respectively. Moreover, 43 drugs are predicted ranked in at the 1 top in HNRD.

D. HNRD PREDICTS NOVEL RD

After confirming the prediction ability of HNDR by cross-validation experiments, we conducted a comprehensive prediction of novel associations between all drugs and diseases. In the inference process, all known drug-disease associations in the gold standard dataset are used as the training set and the remaining drug disease pairs are regarded as the set of candidate drug-disease associations. HNRD can predict the potential disease associations for all drugs simultaneously. By applying HNRD, all candidate diseases for a specific drug are ranked according to their predicted values assigned by HNRD. We also have conducted case studies to verify whether the predicted top-ranked diseases are true or not according to two public biological databases: KEGG [50] and CTD [51], which have been constantly updated to include newly verified drug-disease associations and provide a foundation for our validation. We examined the most potential indications for each of the 593 drugs. The predicted results by all methods are summarized in Supplementary form S1. One can observe that 160 of top-5 predicted novel drug-disease associations by HNRD have been annotated in KEGG and CTD, respectively, which are more than the other prediction methods. We choose several drugs as examples and list the verified information of the top-5 candidate diseases for each selected drug in Supplementary Tables S2. We find several novel drug disease associations of the top-ranked predictions that have been annotated in KEGG, CTD or the other papers. For example Esophageal cancer is cancer arising from the esophagus–the food pipe that runs between the throat and the stomach [52]. Topotecan is a semi-synthetic derivative of camptothecin. Camptothecin is a natural product extracted from the bark of the tree Camptotheca acuminata. Topoisomerase-I is a nuclear enzyme that relieves torsional strain in DNA by opening single strand breaks [53]. Once topoisomerase-I creates a single strand break, the DNA can rotate in front of the advancing replication fork. In physiological environments, topotecan is in equilibrium with its inactive carboxylate form [54], so it also can be used in Esophageal cancer by the same function.

E. VALIDATION ON THE OTHER DATASETS

To demonstrate the capability of HNRD in predicting new drugs related to a queried disease, we also conduct some test on other datasets, including Cdataset and DNdataset, which have been used in the research [27]. Cdataset includes 663 drugs registered in DrugBank, 409 diseases listed in OMIM database, and 2,353 verified drug-disease associations. DNdataset contains 4,516 diseases annotated by Disease Ontology (DO) terms, 1,490 drugs registered in DrugBank and 1008 known drug-disease associations derived from DrugBank.

We conduct ten times ten-fold cross-validation to validate the prediction accuracy of our proposed method on Cdataset and DNdataset. HNRD achieves an AUC value of 0.95 in the Ddataset whereas DRRS, MBiRW, HGBI, DrugNet and KBMF obtain inferior results of 0.947, 0.933, 0.858, 0.804 and 0.928, respectively. The maximum precision achieved by HNRD is 0.67, which is higher than that of the other methods. The AUC value obtained by HNRD is 0.97 in the Dndataset, which is higher than that obtained by DRRS, MBiRW and DrugNet. This may be due to the larger size than the other datasets.

IV. DISCUSSION AND CONCLUSION

Identifying the relationships between drugs and diseases is essential for understanding the mechanisms and functions of drugs. In this paper, we apply a neural-network-based model to predict drug-disease associations. LOOCV and case studies are implemented to evaluate the performance of our method in comparison with the other state-of-the-art approaches. In comparison with the state-of-the-art method, HNRD performs better in terms of AUC values on the dataset and can retrieve more correct associations. Results show that HNRD could be a useful tool for studying the drug-disease relationship. We analyze the top 5 predictions by using HNRD. In the case studies, we confirm drug connections with gastric, ovarian, and colorectal cancer by literature mining. Our study has a major contribution in identifying potential drug-disease associations that our method could integrate more matrix than we have integrated (e.g., drug Gaussian kernel similarity matrix) due to its property of heterogeneity.

The basic idea of considering drug-disease prediction problem as a matrix factorization problem is to determine a low-rank matrix that can integrate prior knowledge about drug and disease. Multiple methods have been proposed and then improved for the task. Therefore, our method might be improved in the future. Considering that matrix factorization is often applied in small data, when the number of data increases, the time consumed is very long. However, HNRD is generated from the neural network, which needs sufficient data. As time passes, the dataset will be updated, and the model will be friendlier to predict.

ACKNOWLEDGMENT

(Yingdong Wang and Gaoshan Deng contributed equally to this work.)

REFERENCES

[1] K. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. L. Barabási, "The human disease network," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 21, pp. 8685–8690, 2007.

[2] L. Weng, L. Zhang, Y. Peng, and R. S. Huang, "Pharmacogenetics and pharmacogenomics: A bridge to individualized cancer therapy," *Pharmacogenomics*, vol. 14, no. 3, pp. 315–324, 2013.

[3] Y. H. Li *et al.*, "Clinical trials, progression-speed differentiating features and swiftness rule of the innovative targets of first-in-class drugs," *Briefings Bioinf.*, 2019.

[4] Y. H. Li *et al.*, "Therapeutic target database update 2018: Enriched resource for facilitating bench-to-clinic research of targeted therapeutics," *Nucleic Acids Res.*, vol. 46, no. D1, p. D1121, 2018.

[5] H. Yang *et al.*, "Therapeutic target database update 2016: Enriched resource for bench to clinical drug target and targeted pathway information," *Nucleic Acids Res.*, vol. 44, no. D1, pp. 1069–1074, 2016.

[6] S. J. Cockell *et al.*, "An integrated dataset for in silico drug discovery," *J. Integr. Bioinf.*, vol. 7, no. 3, pp. 15–27, 2010.

[7] S. Naylor and J. Schonfeld, "Therapeutic drug repurposing, repositioning and rescue—Part I: Overview," *Drug Discovery World*, vol. 16, pp. 49–62, Dec. 2014.

[8] F. Zhu, X. X. Li, S. Y. Yang, and Y. Z. Chen, "Clinical success of drug targets prospectively predicted by in silico study," *Trends Pharmacol. Sci.*, vol. 39, no. 3, pp. 229–231, 2017.

[9] Z.-J. Han, W.-W. Xue, L. Tao, and F. Zhu, "Identification of novel immune-relevant drug target genes for Alzheimer's Disease by combining ontology inference with network analysis," *CNS Neurosci. Therapeutics*, vol. 24, no. 12, pp. 1253–1263, 2018.

[10] Y. Xu, Y. Wang, J. Luo, W. Zhao, and X. Zhou, "Deep learning of the splicing (EPI) genetic code reveals a novel candidate mechanism linking histone modifications to ESC fate decision," *Nucleic Acids Res.*, vol. 45, no. 21, pp. 12100–12112, 2017.

[11] L. Wei, S. Wan, J. Guo, and K. K. L. Wong, "A novel hierarchical selective ensemble classifier with bioinformatics application," *Artif. Intell. Med.*, vol. 83, pp. 82–90, Nov. 2017.

[12] L. Wei, P. Xing, J. Zeng, J. Chen, R. Su, and F. Guo, "Improved prediction of protein–protein interactions using novel negative samples, features, and an ensemble classifier," *Artif. Intell. Med.*, vol. 83, pp. 67–74, Nov. 2017.

[13] Y. Xu, M. Guo, X. Liu, C. Wang, Y. Liu, and G. Liu, "Identify bilayer modules via pseudo-3D clustering: Applications to miRNA-gene bilayer networks," *Nucleic Acids Res.*, vol. 44, no. 20, p. e152, 2016.

[14] Y. Xu, M. Guo, W. Shi, X. Liu, and C. Wang, "A novel insight into gene ontology semantic similarity," *Genomics*, vol. 101, no. 6, pp. 368–375, 2013.

[15] J. Li, S. Zheng, B. Chen, A. J. Butte, S. J. Swamidass, and Z. Lu, "A survey of current trends in computational drug repositioning," *Briefings Bioinf.*, vol. 17, no. 1, pp. 2–12, 2016.

[16] J. Tang *et al.*, "ANPELA: Analysis and performance assessment of the label-free quantification workflow for metaproteomic studies," *Briefings Bioinf.*, 2019.

[17] J. Fu *et al.*, "Discovery of the consistently well-performed analysis chain for swath-ms based pharmacoproteomic quantification," *Frontiers Pharmacol.*, vol. 9, p. 681, Jun. 2018.

[18] B. Li *et al.*, "NOREVA: Normalization and evaluation of MS-based metabolomics data," *Nucleic Acids Res.*, vol. 45, no. W1, pp. W162–W170, 2017.

[19] B. Li *et al.*, "Performance evaluation and online realization of data-driven normalization methods used in LC/MS based untargeted metabolomics analysis," *Sci. Rep.*, vol. 6, Dec. 2016, Art. no. 38881.

[20] N. Zeng, H. Qiu, Z. Wang, W. Liu, H. Zhang, and Y. Li, "A new switching-delayed-PSO-based optimized SVM algorithm for diagnosis of Alzheimer's disease," *Neurocomputing*, vol. 320, pp. 195–202, Dec. 2018.

[21] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, "PREDICT: A method for inferring novel drug indications with application to personalized medicine," *Mol. Syst. Biol.*, vol. 7, no. 1, p. 496, 2014.

[22] H. Luo, M. Li, S. Wang, Q. Liu, Y. Li, and J. Wang, "Computational drug repositioning using low-rank matrix approximation and randomized algorithms," *Bioinformatics*, vol. 34, no. 11, pp. 1904–1912, 2018.

[23] F. Wan, L. Hong, A. Xiao, T. Jiang, and J. Zeng, "NeoDTI: Neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions," *Bioinformatics*, vol. 35, no. 1, pp. 104–111, 2018.

[24] W. Wang, S. Yang, X. Zhang, and J. Li, "Drug repositioning by integrating target information through a heterogeneous network model," *Bioinformatics*, vol. 30, no. 20, pp. 2923–2930, 2014.

[25] V. Martinez, C. Navarro, C. Cano, W. Fajardo, and A. Blanco, "DrugNet: Network-based drug–disease prioritization by integrating heterogeneous data," *Artif. Intell. Med.*, vol. 63, no. 1, pp. 41–49, 2015.

[26] N. Natarajan and I. S. Dhillon, "Inductive matrix completion for predicting gene–disease associations," *Bioinformatics*, vol. 30, no. 12, pp. i60–i68, 2014.

[27] H. Luo *et al.*, "Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm," *Bioinformatics*, vol. 32, no. 17, pp. 2664–2671, 2016.

[28] L. Jiang, Y. Ding, J. Tang, and F. Guo, "MDA-SKF: Similarity kernel fusion for accurately discovering miRNA-disease association," *Frontiers Genet.*, vol. 9, p. 618, 2018.

[29] X. Zeng, L. Liu, L. Lü, and Q. Zou, "Prediction of potential disease-associated microRNAs using structural perturbation method," *Bioinformatics*, vol. 34, no. 14, pp. 2425–2432, 2018.

[30] X. Zhang, Q. Zou, A. Rodriguez-Paton, and X. Zeng, "Meta-path methods for prioritizing candidate disease miRNAs," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 16, no. 1, pp. 283–291, Jan. 2019.

[31] Y. Ding, J. Tang, and F. Guo, "Identification of drug-side effect association via multiple information integration with centered kernel alignment," *Neurocomputing*, vol. 325, pp. 211–224, Jan. 2019.

[32] Y. Liu, X. Zeng, Z. He, and Q. Zou, "Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 4, pp. 905–915, Jul. 2017.

[33] X. Zeng, X. Zhang, Y. Liao, and L. Pan, "Prediction and validation of association between microRNAs and diseases by multipath methods," *Biochim. Biophys. Acta (BBA)-Gen. Subjects*, vol. 1860, no. 11, pp. 2735–2739, 2016.

[34] L. Jiang, Y. Xiao, Y. Ding, J. Tang, and F. Guo, "FKL-Spa-LapRLS: An accurate method for identifying human microRNA-disease association," *BMC Genomics*, vol. 19, no. 10, p. 911, 2018.

[35] C. Long, W. Li, P. Liang, S. Liu, and Y. Zuo, "Transcriptome comparisons of multi-species identify differential genome activation of mammals embryogenesis," *IEEE Access*, vol. 7, pp. 7794–7802, 2019.

[36] Y. Ding, J. Tang, and F. Guo, "Identification of drug-target interactions via multiple information integration," *Inf. Sci.*, vols. 418–419, pp. 546–560, Dec. 2017.

[37] C. Shen, Y. Ding, J. Tang, X. Xu, and F. Guo, "An ameliorated prediction of drug–target interactions based on multi-scale discrete wavelet transform and network features," *Int. J. Mol. Sci.*, vol. 18, no. 8, p. 1781, 2017.

[38] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *J. Chem. Inf. Comput. Sci.*, vol. 28, no. 1, pp. 31–36, 1988.

[39] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. L. Willighagen, "The chemistry development kit (CDK): An open-source Java library for chemo- and bioinformatics," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 2, pp. 493–500, 2003.

[40] M. A. Van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, and J. A. M. Leunissen, "A text-mining analysis of the human phenome," *Eur. J. Hum. Genet.*, vol. 14, no. 5, pp. 535–542, 2006.

[41] A. Hamosh, A. F. Scott, J. S. Amberger, D. Valle, and V. A. Mckusick, "Online mendelian inheritance in man (OMIM)," *Hum. Mutation*, vol. 15, no. 1, pp. 57–61, 2000.

[42] D. S. Wishart *et al.*, "DrugBank: A comprehensive resource for *in silico* drug discovery and exploration," *Nucleic Acids Res.*, vol. 34, no. 1, pp. 668–672, 2006.

[43] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.

[44] Y. Luo *et al.*, "A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information," *Nature Commun.*, vol. 8, no. 1, p. 573, 2017.

[45] F.-Y. Dao *et al.*, "Identify origin of replication in saccharomyces cerevisiae using two-step feature selection technique," *Bioinformatics*, 2018.

[46] Q. Zou, J. Li, L. Song, X. Zeng, and G. Wang, "Similarity computation strategies in the microRNA-disease network: A survey," *Briefings Functional Genomics*, vol. 15, no. 1, pp. 55–64, 2015.

[47] C.-Q. Feng *et al.*, "iTerm-PseKNC: A sequence-based tool for predicting bacterial transcriptional terminators," *Bioinformatics*, 2018.

[48] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proc. Int. Conf. Mach. Learn.*, 2006, pp. 233–240.

[49] M. Gonen and S. Kaski, "Kernelized Bayesian matrix factorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 2047–2060, Oct. 2014.

[50] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Data, information, knowledge and principle: Back to metabolism in KEGG," *Nucleic Acids Res.*, vol. 42, no. 1, pp. 199–205, 2014.

[51] A. P. Davis *et al.*, "The comparative toxicogenomics database: Update 2011," *Nucleic Acids Res.*, vol. 39, no. 1, pp. 1067–1072, 2011.

[52] B. W. Stewart and C. P. Montgomery, "Oesophageal cancer," World Cancer Rep., 2014, pp. 528–543.

[53] Y. Pommier, E. Leo, H. L. Zhang, and C. Marchand, "Dna topoisomerases and their poisoning by anticancer and antibacterial drugs," *Chem. Biol.*, vol. 17, no. 5, pp. 421–433, 2010.

[54] G. Cordell, "The alkaloids: Chemistry and biology," in *The Alkaloids: Chemistry and Biology*, vol. 17, no. 5. New York, NY, USA: Academic, 2003, pp. 1–50.

YINGDONG WANG (S'17–M'17) received the master's degree in educational technology from Sun Yat-sen University, in 2012. She is currently pursuing the Ph.D. degree in computer science with Xiamen University. Her current research interests include big data, human–machine interface, and bioinformatics.

GAOSHAN DENG received the bachelor's degree in software engineering from Xiamen University, in 2017. He is currently pursuing the M.S degree from the Computer Science Department, University of Southern California. His current research interests include multi-objective optimization, big data, and data mining. He is a Student Member the IEEE Computational Intelligence Society.

NIANYIN ZENG received the B.Eng. degree in electrical engineering and automation and the Ph.D. degree in electrical engineering from Fuzhou University, in 2008 and 2013, respectively. From 2012 to 2013, he was a RA with the Department of Electrical and Electronic Engineering, The University of Hong Kong. From 2017 to 2018, he was an ISEF Fellow founded by the Korea Foundation for Advance Studies and also a Visiting Professor with the Korea Advanced Institute of Science and Technology.

He is currently an Associate Professor with the Department of Instrumental & Electrical Engineering, Xiamen University. He has authored or coauthored several technical papers, including six ESI Highly Cited Papers according to the most recent Clarivate Analytics ESI report and also a very active reviewer for many international journals and conferences. His current research interests include intelligent data analysis, computational intelligent, and time-series modeling and applications.

Dr. Zeng is currently serving as an Associate Editor for *Neurocomputing*, an Editorial Board members for *Computers in Biology and Medicine*, *Biomedical Engineering Online*, and also a Guest Editor for *Frontiers in Neuroscience*. He also serves as a Technical Program Committee Member for ICBEB 2014 and an Invited Session Chair of ICCSE 2017.

XIAO SONG received the B.Sc. degree in computer science and technology from the Zhengzhou University of Light Industry, in 2007, and the Ph.D. degree in electrical engineering from Xiamen University, in 2012. From 2014 to 2017, she held a postdoctoral position with the Department of Computer Science and Technology, Huazhong University of Science and Technology. She is currently an Associate Professor with the School of Computer and Information Technology, Nanyang Normal University. Her current research interests include computational intelligent and bioinformatics.

YUANYING ZHUANG received the B.Sc. degree in mathematics and applied mathematics from Xiamen University, in 2007, and the M.Sc. degree in mathematics and computing for finance and the Ph.D. degree in mathematics from Swansea University, U.K., in 2009 and 2013, respectively. From 2009 to 2013, he was a fixed term Tutor with the International College of Wales Swansea. He is currently a Lecturer with the School of Mathematics and Statistics, Nanyang Institute of Technology. His current research interests include stochastic analysis, statistical forecasting, and bioinformatics.