A Drug Repositioning Algorithm Based on Deep Auto-Encoder and Adaptive Fusion

Peng Chen
China Three Gorges University
Tianjiazhi Bao
China Three Gorges University
Xiaosheng Yu
China Three Gorges University
Zhongtu Liu (liuzhongtu@ctgu.edu.cn)
China Three Gorges University

Research article

Keywords: drug repositioning, adaptive fusion, deep auto-encoder

Posted Date: September 8th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-56650/v1

A Drug Repositioning Algorithm based on Deep Auto-Encoder and Adaptive Fusion

Peng Chen Tianjiazhi Bao Xiaosheng Yu Zhongtu Liu*

chenpeng@ctgu.edu.cn baotianjiazhi@outlook.com yuxiaosheng@ctgu.edu.cn liuzhongtu@ctgu.edu.cn

Abstract

Background: Drug repositioning has aroused extensive attention by scholars at home and abroad due to its effective reduction in development cost and time of new drugs. However, the current drug repositioning based on computational analysis methods is still limited by the problems of data sparse and fusion methods, so we use autoencoders and adaptive fusion methods to calculate drug repositioning.

Results: In this paper, a drug repositioning algorithm based on deep auto-encoder and adaptive fusion has been proposed against the problems of declined precision and low-efficiency multi-source data fusion caused by data sparseness. Specifically, the drug is repositioned through fusing drug-disease association, drug target protein, drug chemical structure and drug side effects. To begin with, drug feature data integrated by drug target protein and chemical structure were processed with dimension reduction via a deep auto-encoder, to obtain feature representation more densely and abstractly. On this basis, disease similarity was computed by the drug-disease association data, while drug similarity was calculated by drug feature and drug-side effect data. Besides, the predictions of drug-disease associations were calculated using a Top-k neighbor method that is more suitable for drug repositioning. Finally, a predicted matrix for drug-disease associations has been acquired upon fusing a wide variety of data via adaptive fusion. According to the experimental results, the proposed algorithm has higher precision and recall rate in comparison to DRCFFS, SLAMS and BADR algorithms that use the same data set for computation.

Conclusion: our proposed algorithm contributes to studying novel uses of drugs, as can be seen from the case analysis of Alzheimer's disease. Therefore, it can provide a certain auxiliary effect for clinical trials of drug repositioning

Keywords: drug repositioning; adaptive fusion; deep auto-encoder

1. Background

Drug repositioning, also known as "conventional drug in new use", is to search new uses for drugs with existing indications. Conventionally, the development of new drugs is an expensive and inefficient process, requiring 10 to 15 years for launching a new drug on the market[1]. Targeted at the difficulty in conventionally developing new drugs, the computational analysis method of drug repositioning can offer a new way to develop new drugs, which has grown into a much-talked-about topic for researchers at home and abroad in recent years.

The research object of drug repositioning was selected from the list of drugs approved by the FDA (US Food and Drug Administration). By doing so, the costs and risks can be decreased in comparison to studying unknown new drugs. Developing new drugs in a conventional way involves

* Correspondence: liuzhongtu@ctgu.edu.cn

College of Computer and Information Technology, China Three Gorges University, Hubei, China

Full list of author information is available at the end of the article

detection and clinical trials at early stages, safety review, clinical research, and post-marketing safety monitoring, whereas the drug repositioning method merely needs to go through compound identification, compound acquisition, drug development, and post-marketing FDA safety monitoring. In contrast, there has been an increasing number of successful cases of drug repositioning. For instance, the original research goal of sildenafil was to treat vascular diseases such as angina pectoris; however, its effect on treating male erectile dysfunction was unexpectedly discovered in the course of clinical testing[2]. Later on, it was found from a follow-up study that low-dose sildenafil can also be used for treating rare pulmonary hypertension[3]. Colesevelam is a bile acid sequestrant with an initial indication of primary hypercholesterolemia, which has been approved as a treatment drug for type 2 diabetes since its effect on controlling blood sugar was proven in an experimental study[4].

Certain occasional cases of drug repositioning could be found in earlier times. However, a well-regulated and repeatable drug repositioning method has been studied by researchers at home and abroad after they discovered the significance of drug repositioning to the development of new drugs thanks to the advancement of science and technology over time. Although the drug repositioning method combing machine learning and deep learning has become a mainstream, problems such as ineffective fusion of multi-source data and sparse data remain an obstacle in the study of drug repositioning. Cheng et al.[5] proposed and designed a complete, network-based drug repositioning method to determine new drug targets and anti-cancer indications through labeling significant mutant genes found in the human cancer genome. Zhang et al.[6] computed the drug-disease prediction upon measuring drug similarity through various data and inferred effective drug repositioning results via the fusion of all predictions with the maximum margin method interval method. And the AUC (Area Under Curve) reached up to 0.8949. However, they failed to calculate the drug repositioning results from the disease-oriented perspective due to a single point of consideration. Zhang Jia et al.[7] computed the similarity of multiple data using the improved collaborative filtering equation and fused prediction results with an adaptive method. However, the problem of data sparseness remained. Luo et al.[8] proposed to calculate the similarity of drugs and diseases using similar comprehensive methods of measurement; then, a network was constructed via the two similarities and fused into a heterogeneous network of drug-disease interactions. On this basis, a new drug-disease association could be predicted with two random walk models with AUC reaching 0.917. However, multiple data sources cannot be rapidly and effectively fused using this method. Zeng et al.[9] developed a technique called deepDR to extract drug features from the heterogeneous network adopting multi-modal auto-encoders. And drugs that can be repositioned were inferred through encoding and decoding drugs with low dimension by means of the variation auto-encoder. In that case, AUC reached 0.908, which was superior to the conventional network-based drug repositioning algorithm. In response to the global outbreak of 2019-nCov, Zhou [10] proposed an antiviral drug reuse methodology and determined 135 kinds of drugs for the prevention and treatment of HCoV in accordance with network medical platform based on system pharmacology.

DRDA (Drug Repositioning Algorithm based on Deep auto-encoder and Adaptive fusion has been put forward in this paper to deal with problems of data sparseness and low-efficient fusion of multi-source data in drug repositioning. In response to the sparseness of drug chemical structure and drug target protein data, two types of data were firstly integrated into DRDA. The integrated data are known as "drug feature" data. Subsequently, more abstract features can be captured through

dimension reduction in drug feature data via the deep auto-encoder. Upon dimension reduction in features, disease similarity was computed by the drug-disease associated data, while drug similarity was computed by drug feature data and drug side-effect data. In addition, the predictions of drug-disease associations were calculated using a Top-k neighbor that is more suitable for drug repositioning. Finally, the weight of various data sources was adjusted via adaptive fusion with an aim at decreasing the weight of data sources that cannot provide effective information for drug repositioning. The experimental results show that the designed drug repositioning algorithm can effectively reduce data sparseness and adjust the weight of data source for improving the precision and recall rate. Main contributions are presented as follows:

1. A general drug repositioning framework fusing information of multiple data sources was established to effectively perform drug repositioning computation.

2. A deep auto-encoder was designed for performing dimension reduction on drug feature data (drug chemical structure and drug target protein), which can extract abstract features. In this way, it can not only alleviate the problem of data sparseness but also improve computation efficiency.

3. A weight computation method that is more appropriate for drug repositioning was designed for multi-source data fusion, not only guaranteeing the effect of algorithm indicators but also enhancing the prediction ability of new uses of drugs.

4. The experimental results show that DRDA has an excellent performance in all indicators (precision, recall rate and F-score, etc.). In particular, its precision reaches 0.9047 with the F-score of 0.9041. In the meantime, a better prediction result can be witnessed in drug-disease association because of the particularity of the fusion method.

After that, this paper is structured as follows. Chapter 2 introduces the basic concepts of an encoder. And DRDA is proposed in Chapter 3. Next, DRDA is analyzed and verified via various sets of experiments in Chapter 4. Meanwhile, related cases are illustrated. In Chapter 5, the work that has been conducted ,and future research directions are concluded.

2. Auto-encoder

An auto-encoder is a fully connected neural network with a single hidden layer for performing dimension reduction on data through reconstructing input[11], that is, the output data are set to be equal to the input data. On this basis, data of the hidden layer data extracted can be used as data upon dimension reduction. Its structure is shown in Fig. 1.

Regarding the input $x \in R^{n\times 1}$, the hidden layer features of the input features can be obtained by the encoder. And then, the hidden layer features are reconstructed by the decoder. Besides, the network is optimized through minimizing the error between the input and the reconstructed output [12]. The hidden layer features can be computed in Eq. (1):

$$g(x) = f(\omega x + b) \tag{1}$$

The output features are computed in Eq. (2):

$$x' = f(\omega' g(x) + b') \tag{2}$$

Where ω, b, ω', b' are the weight and offset of the encoder and the decoder; and $f(g)$ represents the non-linear activation function. The loss function of the auto-encoder is shown in Eq. (3):

$$J(\omega, b) = \frac{1}{2} \| x' - x \|^2 + \delta \| \omega \|_2^2 \tag{3}$$

Where δ is a hyperparameter that is used for controlling the size of the parameter ω. The

output x' can be approximately equal to the input x by means of minimizing the loss function, of which the hidden layer features can be deemed as low dimension and high level of abstract features of the input features.

3. Methods

DRDA (Drug Repositioning Algorithm based on Deep auto-encoder and Adaptive fusion) has been put forward in this paper to deal with problems of data sparseness and low-efficient fusion of multi-source data in drug repositioning. Firstly, dimension reduction was performed on drug features (incl. drug chemical structure and drug target protein) using the deep auto-encoder before extracting more abstract representations of drug features. Then, drug similarity was computed in line with drug features and drug side-effect data, while disease similarity was computed using drug-disease associated data. Moreover, the prediction of the drug-disease association was computed using the Top-k similar neighbor method that is more suitable for drug repositioning. Finally, the prediction of the drug-disease association was acquired with the fusion of predictions computed by various data sources utilizing adaptive fusion. The algorithm framework is shown in Fig. 2.

3.1 Dimension reduction in drug features based on the deep auto-encoder

In order to cut down the sparseness of drug chemical structure and drug target protein data, two types of data were firstly integrated into drug feature data. After that, more abstract drug features were extracted to lower the sparseness of data through performing dimension reduction on drug features with a deep auto-encoder. The deep auto-encoder is an extension of the auto-encoder, which converts high-dimensional data to low-dimensional data via a multi-layer encoding network and recover the encoding with a similar decoder. And a training network for errors between input data can be reconstructed by minimizing the original data[13]. On this basis, a deep auto-encoder that is more relevant to drug feature data was designed. Its structure is shown in Fig. 3.

The deep auto-encoder is composed of an encoder and a decoder. And "Adagrad" [14] is regarded as an optimization method. The encoder consists of an input layer and three building layers [15]. To be specific, the building layer is composed of a fully connected layer and a discarded layer, and the last building layer is a coding layer. And ReLU is used as the activation function. When dimension reduction is performed on drug feature data, with drug feature data $x \in R^{m \times n}$ as input, the building layer is computed in Eq. (4):

$$g^i(x) = f(\omega^i g^*(x) + b^i) \tag{4}$$

Where, $g^i(x)$ is the output of the building layer; $f(\mathrm{g})$ is a nonlinear activation function; ω^i, b^i represents the weight and offset of the ith building layer, $g^*(x) = \begin{cases} x & i = 1 \\ g^{i-1}(x) & i > 1 \end{cases}$; and i presents the number of building layers.

The decoder is comprised of three building layers and an output layer (the first building layer is an encoding layer). ReLu is regarded as the activation function in all layers except the output layer

that uses Sigmoid as its activation function. As the drug feature data are binary data, the reconstructed output should be data approaching 0 or 1 in order to reduce errors between the reconstructed output and input as quickly as possible. Most of the output values of the Sigmoid function are concentrated around 0 and 1 (as shown in Fig. 4), which is more in line with the structure required by the output. The computation method of the decoder building layer is similar to that of the encoder building layer. The decoder output layer can be computed in Eq. (5):

$$\bar{x} = g^4(x) = f(\omega^3 g^3(x) + b^3) \tag{5}$$

Mean square error is applied in the deep auto-encoder as the loss function. The gap between the input x and the output $\bar{x}$ upon reconstructed input can be decreased by minimizing the mean square error. With the minimal error, extracting the features of the encoding layer is the feature data after dimension reduction.

The training parameter batch size adopted is 16, and the learning rate of the optimization function "Adagrad" is 0.01. Before training, parameters are initialized via Xavier. And 400-dimensional feature data extracted from the output of the encoding layer are used for subsequent computation of drug similarity.

3.2 Similarity computation

3.2.1 Drug similarity

Dudley[16], Li[17] et al. believed that drug chemical structure and drug target protein play a crucial role in calculating drug similarity since they are quantitatively related. Moreover, drugs with similar target proteins can also treat similar diseases. Drug chemical structure data and drug target protein data were sampled from PubChem[18] and UniPort Knowledgebase[19]. After dimension reduction of drug feature data, the drug features that are denser than the original data can be obtained. Then, cosine similarity is used for calculating drug similarity, as shown in Eq. (6):

$$sim(d, d^*) = \frac{\sum_{i=1}^{n} \mathrm{f}_{d_i} \bullet \mathrm{f}_{d_i^*}}{\sqrt{\sum_{i=1}^{n} \mathrm{f}_{d_i}^2} \bullet \sqrt{\sum_{i=1}^{n} \mathrm{f}_{d_i^*}^2}} \tag{6}$$

$sim(d, d^*)$ represents the similarity of the drug d and the drug d^*; f_{d_i} and $\mathrm{f}_{d_i^*}$ stand for the value of the ith drug feature in the drug d and the drug d^*, respectively; and n is feature dimension.

It was proposed to verify whether the two drugs can act on the same target through the drug side effect data in literature[20]. Also, a series of experiments were designed to prove the feasibility of inferring the molecular interaction with the side effect data. Thus, the drug side effect data can be used for calculating drug similarity. The drug side effect data were sampled from the SIDER[21] database. If the drug causes this type of side effect, the value will be set as 1; if not, it will be as 0. Tanimoto Coefficient is used for computation in Eq. (7):

$$sim(d, d^*) = \frac{|I_{dd^*}|}{|I_d| + |I_{d^*}| - |I_{dd^*}|} \tag{7}$$

Where I_{dd*} presents the number of same side effects of the two drugs; I_d and I_{d^*} represents the number of side effects of the two drugs, respectively.

3.2.2 Disease similarity

In literature [22], an inferred idea associated with drug repositioning has been put forward, that is, two diseases are deemed as similar when they can be treated by a variety of identical drugs. According to the method proposed in the literature [7], disease similarity can be computed with drug-disease associated data sampled from UMLS[23]. The data are binary data; namely, if the drug has a treatment effect on the disease, it is deemed as 1; otherwise, it is 0. When Tanimoto Coefficient is used for computation, it is presented as Eq. (8):

$$sim(e, e^*) = \frac{|I_{ee^*}|}{|I_e| + |I_{e^*}| - |I_{ee^*}|} \tag{8}$$

Where, $sim(e, e^*)$ indicates the similarity of the two diseases; I_{ee^*} the number of drugs that can treat the two diseases; I_e and I_{e^*} represent the number of drugs that can treat the diseases e and e^*, respectively.

3.3 Computation of prediction

Only "0" and "1" relationship between the drug and the disease can be found in the original drug-disease associated data, whereas certain side-effect relationship might be detected between the drug and some diseases. In order to effectively calculate the drug-disease associated prediction, the known drug-side effect relationship in the drug-side effect data was marked in the drug-disease data, that is, if a side effect (disease) exists in both the drug-disease associated data and drug-side effect data, the corresponding drug-side effect (disease) should be changed from "0" to "-1" in the drug-disease associated data when the drug produces the side effect.

As can be seen from the literature [7] and [24], the prediction of drug-disease association in drug repositioning was computed using collaborative filtering. Nevertheless, the drug-disease prediction of the conventional approaches of collaborative filtering and Top-k neighbor cannot be accurately computed since data used in drug repositioning was sparse, and there is a side effect relationship between drugs and diseases. In this paper, the drugs or diseases to be evaluated were also computed as similar neighbors on the basis of Top-k proximity. With a small number of effective neighbors (the similarity is not 0) caused by data sparseness, the drug-disease associated information of the drug or disease is decisive, which can avoid false predictions resulted from a lack of effective neighbors. Meanwhile, when there are enough effective neighbors, the drug-disease associated information of the drug or disease is only one of the neighbors, exerting a small effect on predicting the new effect of the drug. Fusing own information for the prediction can lower the prediction error caused by a lack of effective neighbors due to data sparseness. Moreover, applying own information of drugs or diseases for computation can avoid computation collapse resulted from an effective neighbor being zero.

The drug-disease associated prediction can be computed through drug similarity, as shown in Eq. (9):

$$P_{de}^{k}=\frac{\sum_{d^{*}\in \mathrm{NN}'} sim^{k}(d,d^{*})\times s_{d^{*}e}}{\sum_{d^{*}\in \mathrm{NN}'} sim^{k}(d,d^{*})} \tag{9}$$

Where P_{de}^{k} is the predicted score between the drug *d* and the disease *e* computed based on the data source *k*. NN' is the union set of the drug d and its Top-k neighbors. $s_{d^{*}e}$ is the relational value between the drug d^{*} and the disease e upon integrating drug-side effect data in the drug-disease associated data set.

The computation of the drug-disease associated prediction for disease similarity is similar to that of drug similarity, as shown in Eq. (10):

$$P_{de}^{k}=\frac{\sum_{e^{*}\in \mathrm{NN}'} sim^{k}(e,e^{*})\times s_{de^{*}}}{\sum_{e^{*}\in \mathrm{NN}'} sim^{k}(e,e^{*})} \tag{10}$$

Where, NN' is the union set of disease *e* and its Top-k neighbors.

The drug-disease associated prediction that finally fuses multiple data sources is calculated as Eq. (11):

$$P_{de}^{*}=\sum_{k=1}^{\mathrm{K}}\beta_{k}\times P_{de}^{k} \tag{11}$$

Where, P_{de}^{*} is the prediction of the drug *d* after data fusion to the disease *e*; β_{k} is the weight of the k data source; and P_{de}^{k} is the prediction of the drug *d* to the disease *e* in the k data source.

3.4 Weight computation

A weight computation method that is more suitable for drug repositioning was designed per the literature [7], so as to effectively fuse the predictions computed by multiple data sources. Besides, the best prediction effect was achieved by maximizing the combination of the drug-disease associated value being 0 and the difference between its predictions, while minimizing the difference between the combination of the drug-disease value being 1 and its prediction. The weight computation method can be expressed as an optimization objective function, such as (12):

$$\arg\min_{\beta_{k}} L(\beta_{k})=\sum_{k=1}^{k}\beta_{k}^{2}\left(\sum\nolimits_{\{(d,e)|s_{de}=1\}}(s_{de}-p_{de}^{k})^{2}-\sum\nolimits_{\{(d,e)|s_{de}=0\}}(s_{de}-p_{de}^{k})^{2}\right)$$

$$\text{s.t.}\sum_{k=1}^{K}\beta_{k}=1 \tag{12}$$

$$\text{s.t.}\ \beta_{k}>0$$

Where $\{(d,e)\,|\,s_{de}=1\}$ represents the combination of the associated value being 1 between

the drug *d* and the disease *e* in the drug-disease associated data. Similarly, $\{(d,e)|s_{de}=0\}$ is the drug-disease combination being 0. The optimized problem is solved using the minimize method in Scipy of the Python package.

3.5 Algorithm flow

The overall algorithm flow is presented as follows:

Algorithm Drug Repositioning Algorithm based on Deep Auto-Encoder and Adaptive Fusion

Input: Drug-disease correlation matrix $R^{m\times n}$, drug feature data, drug-side effect data, and number of neighbors λ

1. Perform dimension reduction on drug feature data using the deep auto-encoder
2. Calculate drug similarity with Eq. (6) and Eq. (7), and disease similarity with Eq. (8);
3. Integrate drug-disease associated data and drug-side effect data;
4. **for** k in [1:K]:
5. **for** d in [1:m]:
6. **for** e in [1:n]:
7. Calculate the associated prediction P_{de}^{k} between the drug d and the disease e using neighbor parameters via Eq. (9) and Eq. (10);
8. **end**
9. **end**
10. Output the drug-disease associated prediction matrix Pred^{k} based on the k data source;
11. **End**
12. Calculate the weight β^{k} of fusing multiple data sources in Eq. (12), and acquire the drug-disease associated prediction matrix Pred^{*} through fusing all data source predictions.

Output: Drug-disease associated prediction matrix Pred^{*}, weight vector $\beta=\{\beta^{1},\beta^{2}\ \ldots\ldots,\ \beta^{k}\}$

4. Results

To prove the effectiveness and feasibility of the proposed algorithm, data source comparison, method comparison and case analysis were explained in the experiment.

4.1 Data set

Data set used in the experiment involves drug-disease associated data set, drug feature data set

(drug target protein and drug chemical structure) and drug-side effect data set. Specifically, the drug-disease associated data set covers 536 drugs and 578 diseases (incl. 2229 drugs with known treatment effect-disease association); drug feature data includes 775 target proteins and 881 chemical structures; while drug-side effect data contains 1385 side effects. The sparse degrees of the three data sets (the proportion of invalid data in data) are 0.9928, 0.9231, and 0.9455, respectively. It is worth noting that dimension reduction cannot be performed on the drug-disease associated and drug-side effect data due to their particularity. In that case, dimension reduction was performed on drug features with the deep encoder for extracting more abstract expression. The sparse degree of the drug feature data set upon dimension reduction is 0.7703, which is reduced by 16%, compared with the original data set. The number of diseases that can be treated by each drug in the drug-disease associated data are presented in Fig. 5. Clearly, there are approximately 75% of the drugs that can treat less than 5 types of diseases only.

4.2 Experimental indicators

The drug repositioning task in the experiment can be deemed as a binary classification task [6], that is, the drug is either curable (the value is 1) or incurable (the value is 0) for diseases. Precision, recall rate, F-score and ROC curve are experimental indicators. To begin with, confusion matrix is defined in table 1 before defining the above four indicators. Then, the precision and the recall rate are defined as Eq. (13) and Eq. (14):

$$P = TP / (TP + FP) \tag{13}$$

$$R = TP / (TP + FN) \tag{14}$$

On this basis, F-score can be computed, as shown in Eq. (15):

$$F = 2 \times P \times R / (P + R) \tag{15}$$

In addition, the AUC is adopted as the experimental indicator. Since a specific threshold is required for dividing the predicted score into 1 and 0, the threshold should be set in a way that can achieve the best effect of F-score. 10 times of experiments are conducted using the training set and test set in 9:1. And the average value is taken.

4.3 Comparison of data sources

In order to determine the number of neighbors that can achieve the best algorithm effect, changes of AUC of three types of data under different number of neighbors are given. Besides, computation results with over 50 neighbors are given to ensure that there are enough neighbors to provide information, as shown in Fig. 6. It can be seen from Fig. 6 that AUC values of the three data sources are the highest when the number of neighbors is 50. Therefore, 50 is selected as the number of similar neighbors.

After determining the values of the similar neighbor parameters λ, weight computation is performed in accordance with the method mentioned in section 3.4. The highest weight of the drug-disease associated data source is 0.6605 among the three data sources, and the drug feature data and drug side effect data account for 0.1356 and 0.2037, respectively. What's more, three data sources are computed separately and compared with DRDA over precision, recall rate, F-score and AUC values, as shown in Table 2. As can be seen from data in Table 2, four indicators of the drug-disease associated data are the highest among the three data sources, with a precision rate of 0.9823 and an AUC value of 0.9998. And the recall rate of drug features can reach up to 0.8593, while the precision rate is extremely low, reaching 0.2082. Similarly, the drug-side effect can reach

0.6008 in the recall rate; yet its precision is only 0.3090. Apparently, the prediction computed by a single data source is unstable, which cannot well complete the task of drug repositioning. DRDA fuses the results of the three data sources through an adaptive method, which can obtain results that are superior to the three data sources, presenting the precision of 0.9047 and F-score of 0.9041.

As can be observed from the AUC indicator in Table 2, high AUC values of all data sources do not imply that results of all data sources conform to ideal values since drug repositioning is an extremely unbalanced problem. In the drug-disease associated data, the number of drug-disease association being 1 is far less than the number of drug-disease association being 0. Consequently, it leads to blind optimism of AUC values[25].

4.4 Method comparison

In order to evaluate the performance of DRDA, DRDA was compared with SLAMS[6], DRCFFS[7], and DRBC proposed in the literature[24]. The three algorithms were experimented for 10 times using the training set and test set in 9:1. And the average value is taken to ensure the rationality and fairness of the experiment. The numbers of neighbors of the SLAMS, DRCFFS, and DRBC algorithms are all set as 90. As can be observed from Fig. 7(a), the P-R curve of DRDA wraps the P-R curves of the other three algorithms. ROC curves of four algorithms are displayed in Fig. 7 (b). It can be seen that the AUC values of DRDA and DRCFFS are similar and the highest among the four algorithms, being 0.9993 and 0.9994, respectively. And the AUC value of SLAMS is the lowest, presenting 0.8363. Furthermore, precision, recall rate and F-score of the three algorithms and DRDA were also compared in the experiment, as shown in Table 3. DRDA has superior performances over three indications in comparison with the other three conventional algorithms, reaching the precision rate of 0.9047 and the recall rate of 0.9035. Compared with DRCFFS, the best drug repositioning algorithm at the present stage, although DRDA is slightly lower in AUC, it has better performances on the other three indicators than DRCFFS. In particular, its recall rate is 0.1104 higher than that of DRCFFS. Also, DRDA pays more attention to the prediction of unknown drug-disease associations under the premise of guaranteeing indicators. All in all, DRDA has a better performance than the conventional drug repositioning algorithm.

The drug-disease associated numbers (the associated value of the drug and the disease is 1) correctly predicted by the four algorithms are given at different thresholds, so as to extensively compare the drug-disease associated prediction effects of DRDA, DRCFFS, SLAMS, and DRBC. It can be seen from Fig. 8(a) that the abilities of DRDA and DRCFFS predicting the drug-disease associations with known treatment effects are better than that of the other two algorithms. As the threshold is gradually relaxed, DRDA has a better effect than DRCFFS. The proportions of the total number of drug-disease associations correctly predicted by the four algorithms with known treatment relationships at different thresholds to all drug-disease with known treatment relationships are presented in Fig. 8 (b). It can be seen that DRDA and DRCFFS have nearly predicted all drug-disease combinations with known treatment relationships in the course of selecting drugs with the top 30 predicted scores of each disease. In addition, differing from DRCFFS, DRDA also pays attention to the prediction of unknown drug-disease association.That is, the predicted score of the drug-disease associated value being 0 in the original data set is maximized in Eq. (12), apart from predicting the drug-disease association with known treatment relationships. Hence, DRDA can well predict the drug-disease associations with unknown existing relationships on the premise of ensuring the effective prediction of the number of drug-disease associations with

known treatment relationships.

It can be seen from Table 4 that DRDA is slightly lower than the simple average fusion method over the precision, while but DRDA is 0.2364 higher than the simple average fusion method over the recall rate. Moreover, the simple average fusion method cannot effectively filter the data added subsequently, which is not conducive to the fusion of multi-source data. Without doubt, it can be concluded that DRDA is superior to the simple average fusion method in theory and indicators.

4.5 Case Study

In order to prove that DRDA can effectively assist drug repositioning, the drugs of the top ten predictions for Alzheimer's disease (AD) and their original uses are presented in Table 5. Among them, four drugs have been used for the clinical treatment of Alzheimer's disease, and five out of the remaining six drugs have been studied for the treatment of Alzheimer's disease. However, this does not imply that all drug repositioning results predicted by DRDA are feasible since the development of drugs is a long and complicated process. Apparently, the case analysis showed DRDA could save time for developing new drugs by providing theoretical and data support for drug repositioning and assistance in the research direction.

5. Discussion

In this section, we verify the DRDA algorithm through experiments and case study, and the results fully prove the effectiveness and feasibility of DRDA. In other words, case study shows that most of the drugs predicted by the DRDA algorithm to treat Alzheimer's disease have been studied by pharmacists.

6. Conclusion

A drug repositioning algorithm based on deep auto-encoder and adaptive fusion (DRDA) is proposed. Specifically, dimension reduction was performed on data via the deep auto-encoder to decrease the impact resulted from data sparseness. Meanwhile, weights of all data sources were computed to fuse the information of various data sources. The experimental results show that DRDA is superior to the conventional drug repositioning algorithm, reaching the precision of 0.9047. Besides, a weight computation method that is more suitable for drug repositioning was also designed, which can guarantee the quality of indicators and the prediction drug repositioning on the association of new drugs and diseases. Compared with the simple average fusion method, DRDA can efficiently distinguish valid and invalid data sources, although it has a loss in precision. In other words, DRDA is more aligned with the prediction of drug repositioning.

Although significant achievements have been made in the drug repositioning technology thanks to the efforts of researchers at home and abroad, evaluation indicators that are perfectly correspondent to drug repositioning cannot be found in existing studies, as mentioned in the literature[25]. The drug repositioning algorithm should be evaluated with the combination of related information such as predicted results and side effects, rather than directly considering the predicted new drug-disease result as erroneous. Hence, the performance of the drug repositioning algorithm cannot be accurately determined merely using indicators such as precision, recall rate, and AUC. The follow-up research will focus on enriching and improving existing evaluation indicators, and expanding data sources to achieve accurate disease similarity computation on the basis of the current study.

Abbreviations

AUC: The area under the receiver operating characteristic curve;DRDA: A Drug Repositioning Algorithm based on Deep Auto-Encoder and Adaptive Fusion; DRCFFS: Computational drug repositioning using collaborative filtering by fusing multi-source data; SLAMS:Similarity-based LArge-margin learning of Multiple Sources; DRBC: Drug repositioning algorithm based on collaborative filtering;UMLS: Unified Medical Language System

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All the original data come from the literature[6] of Prof. Zhang. The processed datasets during the current study are available from the author on reasonable request.

Competing interests

The authors declared that they have no competing interests exist.

Funding

The work is supported by The National Key Research and Development Program of China (Grant No.2016YFC0802500) and Sponsored by Research Fund for Excellent Dissertation of China Three Gorges University(Grant No.2020SSPY075)

Authors' contributions

BTJZ wrote the manuscript and developed the source codes.CP and YXS revised the manuscript. LZT collected the datasetsand and contacted relevant units. All authors contributed to the conception and design of the study, participated in the analysis of the results, and edited the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We are very grateful to Prof. Zhang in Watson Research Center for providing data support for this article, and also very grateful to Dr. Zhang from Xiamen University for his comments and suggestions for this article

Author details

College of Computer and Information Technology, China Three Gorges University, Hubei, China

References

[1] Lotfi Shahreza M, Ghadiri N, Mousavi S R, et al. A review of network-based approaches to drug repositioning. Briefings in bioinformatics, 2018, 19(5): 878-892.

[2] Booth B, Zemmel R. Quest for the best. Nat Rev Drug Discov, 2003, 2 (10): 838-841.

[3] Sardana D, Zhu C, Zhang M, et al. Drug repositioning for orphan diseases. Briefings in bioinformatics, 2011, 12(4): 346-356.

[4] Nwose O M, Jones M R. Atypical mechanism of glucose modulation by colesevelam in patients with type 2 diabetes. Clinical Medicine Insights: Endocrinology and Diabetes, 2013, 6(6): 75-79.

[5] Cheng F, Zhao J, Fooksa M, et al. A network-based drug repositioning infrastructure for precision cancer medicine through targeting significantly mutated genes in the human cancer genomes. Journal of the American Medical Informatics Association, 2016, 23(4): 681-691.

[6] Zhang P, Agarwal P, Obradovic Z. Computational drug repositioning by ranking and integrating multiple data sources//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, 2013: 579-594.

[7] Zhang J, Li C, Lin Y, et al. Computational drug repositioning using collaborative filtering via multi-source fusion. Expert Systems with Applications, 2017, 84: 281-289.

[8] Luo H, Wang J, Li M, et al. Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. Bioinformatics, 2016, 32(17): 2664-2671.

[9] Zeng X, Zhu S, Liu X, et al. deepDR: a network-based deep learning approach to in silico drug repositioning. Bioinformatics, 2019, 35(24): 5191-5198.

[10] Zhou Y, Hou Y, Shen J, Huang Y, Martin W, Cheng F. Network-based drug repurposing for human coronavirus. 2020. https://doi.org/10.1101/2020.02.03.20020263.

[11] Alain G, Bengio Y. What regularized auto-encoders learn from the data-generating distribution. The Journal of Machine Learning Research, 2014, 15(1): 3563-3593.

[12] Liang Y,Ji J. Brain functional connections classification method based on prototype learning and deep feature fusion. Acta Automatica Sinica. 2020.https://doi.org/10.16383/j.aas.c190747.

[13] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. science, 2006, 313(5786): 504-507.

[14] Ward R, Wu X, Bottou L. Adagrad stepsizes: Sharp convergence over nonconvex landscapes, from any initialization. arXiv preprint arXiv. 2018.

[15] Moridi M, Ghadirinia M, Sharifi-Zarchi A, et al. The assessment of efficient representation of drug features using deep learning for drug repositioning. BMC bioinformatics, 2019, 20(1): 577.

[16] Dudley J T, Deshpande T, Butte A J. Exploiting Drug-disease relationships for Computational Drug Repositioning . Briefings in Bioinformatics, 2011, 12(4):303-311.

[17] Li J , Zhu X , Chen J Y , et al. Building Disease-Specific Drug-Protein Connectivity Maps from Molecular Interaction Networks and PubMed Abstracts. PLoS Computational Biology, 2009, 5(7):631-641.

[18] Wang Y, Xiao J, Suzek T O, et al. PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Research, 2009, 37(Web Server):W623-W633.

[19] Rolf A , Amos B , Wu C H , et al. UniProt: the Universal Protein knowledgebase. Nucleic Acids Research, 2004, 32(Data Issue):D115-D119.

[20] Campillos M , Kuhn M , Gavin A C , et al. Drug Target Identification Using Side-Effect Similarity. Science, 2008, 321(5886):263-266.

[21] Kuhn M , Campillos M , Letunic I , et al. A side effect resource to capture phenotypic effects of drugs. Molecular Systems Biology, 2010, 6(01): 343-343.

[22] Chiang A P , Butte A J . Systematic Evaluation of Drug–Disease Relationships to Identify Leads for Novel Drug Uses. Clinical Pharmacology & Therapeutics, 2009, 86(5):507-510.

[23] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research, 2004, 32(suppl_1): D267-D70.
[24] Lin Y, Zhang J, Lin M, et al. Drug repositioning algorithm based on collaborative filtering. Journal of Nanjing University(Natural Sciences), 2015, 51(04):834-841.
[25] Brown A S, Patel C J. A standard database for drug repositioning. Scientific data, 2017, 4(01): 1-7.
[26] Choi Y, Jeong H J, Liu Q F, et al. Clozapine improves memory impairment and reduces Aβ level in the Tg-APPswe/PS1dE9 mouse model of Alzheimer's disease. Molecular neurobiology, 2017, 54(01): 450-460.
[27] Zhou Y, Long M, Ran H. Clinical trial of donepezil hydrochloride tablets combined with clozapine dispersible tablets in the treatment of elderly Alzheimer's disease with behavioral disorder. The Chinese Journal of Clinical Pharmacology, 2017, 33(19):1897-1899.
[28] Bennett J, Burns J, Welch P, et al. Safety and tolerability of R (+) pramipexole in mild-to-moderate Alzheimer's disease[J]. Journal of Alzheimer's Disease, 2016, 49(4): 1179-1187.
[29] Wu Y, Xu W, Liu X, et al. Efficacy and Safety of Olanzapine Combined with Repetitive Transcranial Magnetic Stimulation in the Treatment of Behavioral and Psychological Symptoms of Alzheimer's Desease. Herald of Medicine, 2016, 35(10):1069-1072.
[30] Zhang L, Wang L, Wang R, et al. Evaluating the effectiveness of GTM-1, rapamycin, and carbamazepine on autophagy and Alzheimer disease. Medical science monitor: international medical journal of experimental and clinical research, 2017, 23: 801-808.
[31] Sorial M E, El Sayed N S E D. Protective effect of valproic acid in streptozotocin-induced sporadic Alzheimer's disease mouse model: possible involvement of the cholinergic system. Naunyn-Schmiedeberg's archives of pharmacology, 2017, 390(6): 581-593.

Figure Legends

Figure 1. Structure diagram of Auto-encoder

Figure 2. Framework of the drug repositioning algorithm based on deep auto-encoder and adaptive fusion. It includes four types of origin data and three types of data for experiments.

Figure 3. A schematic diagram of a deep auto-encoder framework for drug features data dimensionality reduction. There are two sections to the overall structure: one is the encoder and the other is the decoder. The encoder contains four layers: an input layer, two building layers and an encoding layer; the decoder includes four layers: an encoding layer, two building layers, and an output layer. The building layer is composed of a Dense and Dropout.

Figure 4. Sigmoid function image

Figure 5. The number of diseases that can be treated by each drug in the drug-disease associated data

Figure 6. The figure provides the AUC value changes of the three data sources under different neighbors. In order to ensure the number of effective neighbors, we respectively give the AUC values under the number of neighbors from 50 to 120. Through comparison, it is found that the AUC value of the three data sources is the highest under the number of 50 neighbors.

Figure 7. Figure (a) provides a the PR curves of the DRDA algorithm and the other three algorithms. The PR curve of DRDA wraps the PR curves of the other three algorithms, so the effect is the best; Figure (b) shows the ROC of the four algorithms Comparing the curve and AUC value, DADR and DRCFFS have the best effect.

Figure 8. Figure (a) provides correct prediction of the drug-disease associated number with known treatment relationships for the four algorithms at different thresholds. Figure (b) provides the

proportion of the number of drug-disease associations correctly predicted by the four algorithms with known treatment relationships at different thresholds to all drug-disease with known treatment relationships

Figures and tables

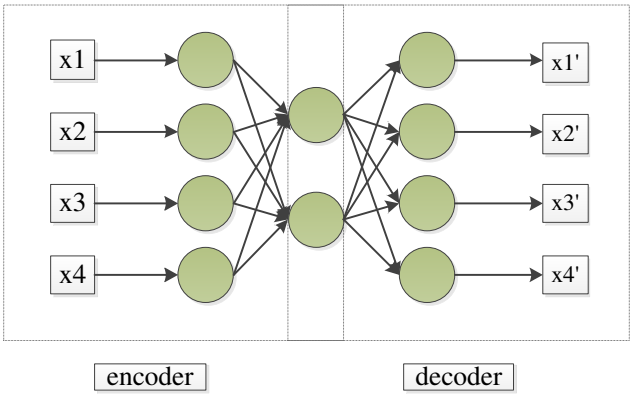


Fig. 1 Auto-encoder

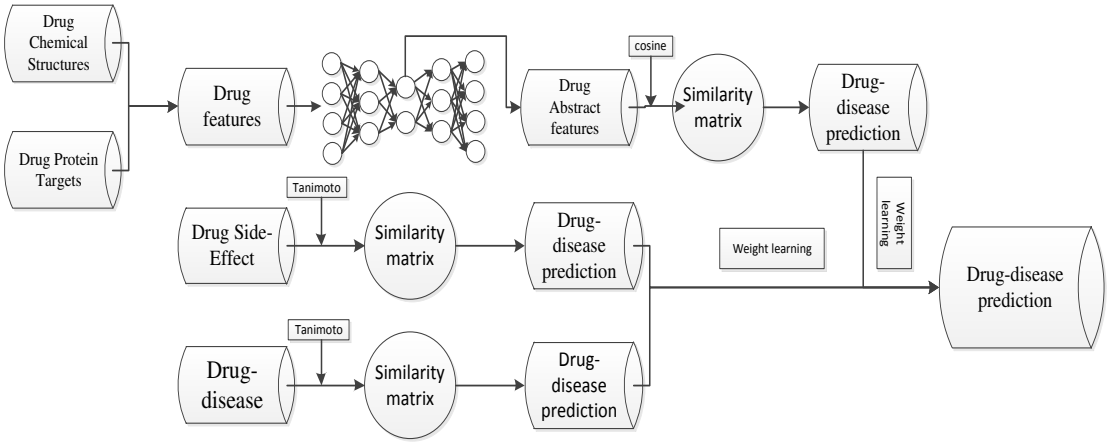


Fig. 2 Framework of the drug repositioning algorithm based on deep auto-encoder and adaptive fusion

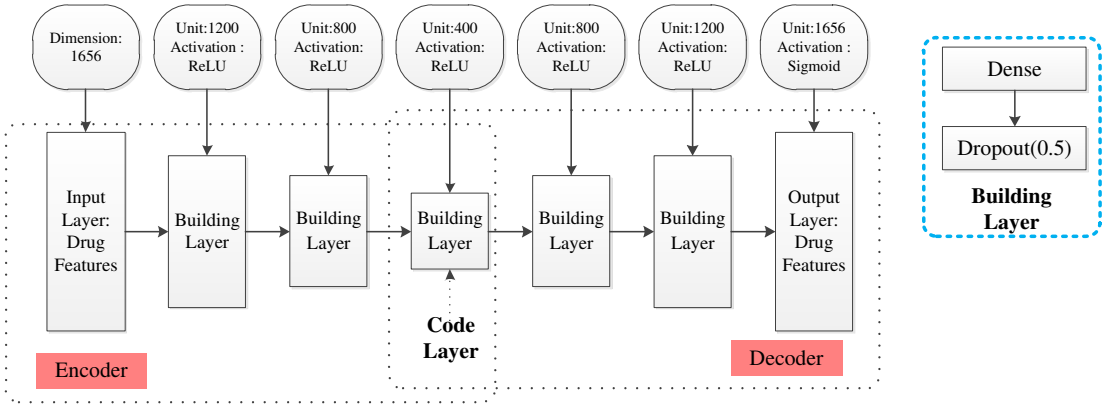


Fig.3. The structure of the deep auto-encoder used for dimension reduction of drug feature data

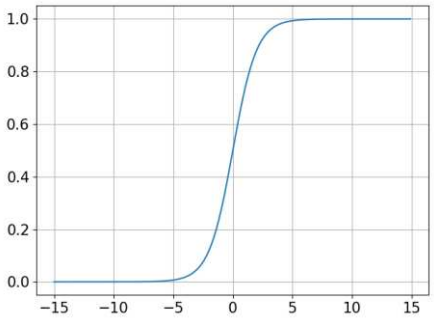


Fig. 4 Sigmoid function image

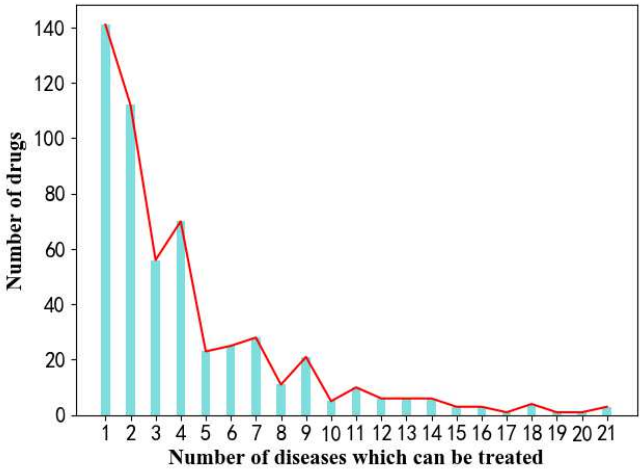


Fig. 5 Distribution of the number of diseases treated by drugs

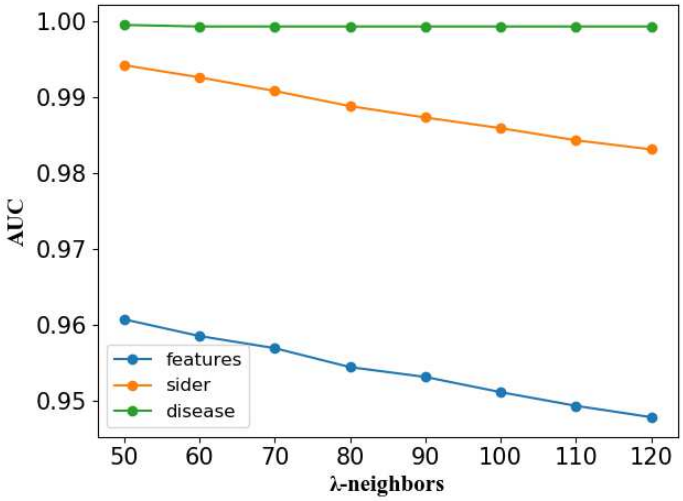
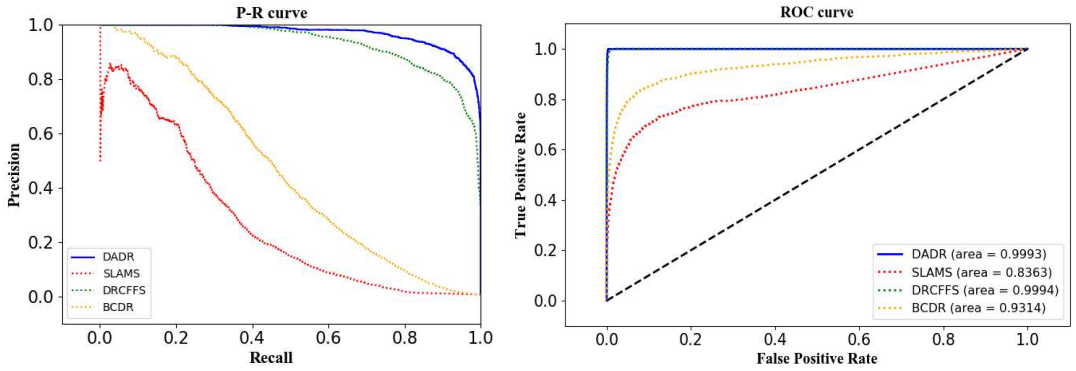
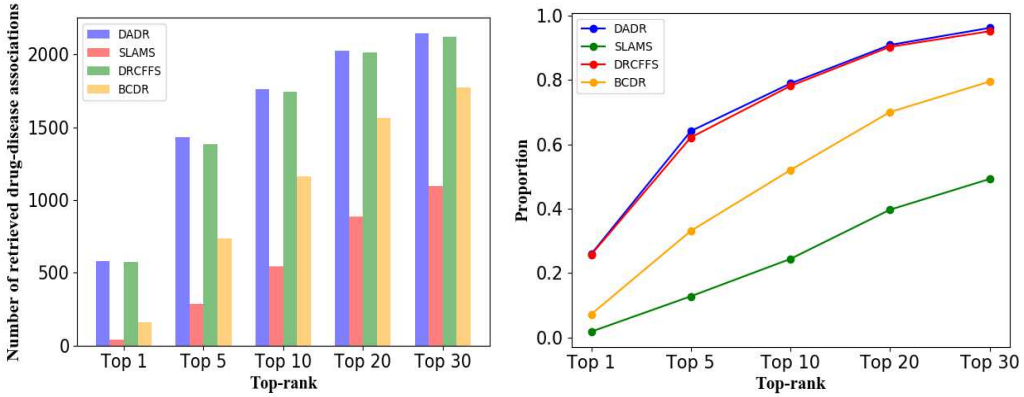


Fig. 6. AUC values of various data sources under different numbers of neighbors



（a） (b)

Fig. 7 Comparison of the P-R curves and ROC curves of the four algorithms



(a) (b)

Fig. 8 Comparison of the four algorithms correctly predicting drug-disease association with known treatment relationships

Table 1 Confusion matrix

	Prediction	
True value	TP (True Positive)	FN (False Negative)
	FP(False Positive)	TN (True Negative)

Table 2 Comparison of three data sources and three indicators of DRDA

Datasets	Precision	Recall	F-score	AUC
Drug-disease	0.9823	0.7167	0.8287	0.9998
Drug feature	0.2082	0.8593	0.3351	0.9649
Drug-side	0.3090	0.6008	0.4081	0.9940
DRDA	0.9047	0.9035	0.9041	0.9993

Table 3 Comparison of precision, recall rate and F-score indicators of the four algorithms

Algorithm	Precision	Recall	F-score
SLAMS	0.1494	0.4984	0.2299
DRCFFS	0.8778	0.7931	0.8333
DRBC	0.7367	0.2768	0.4024
DRDA	0.9047	0.9035	0.9041

Table 4 Comparison of DRDA and the simple average fusion method over precision, recall rate and F-score

Algorithm	Precision	Recall	F-score
DRDA	0.9047	0.9035	0.9041
Simple Avg	0.9441	0.6671	0.7818

Table 5 Drugs with top 10 scores for treating Alzheimer's disease

Drug Name	Origin Use	Study on the treatment of AD
Clozapine	Schizophrenia	[26], [27]
Pramipexole	Parkinsonism	[28]
Olanzapine	Schizophrenia	[29]
Carbamazepine	Epilepsy Peripheral neuralgia	[30]
Donepezil	AD	Used in clinical treatment of Alzheimer's disease
Ethosuximide	Clonic	No research
Galantamine	TO	Used in clinical treatment of Alzheimer's disease
Rivastigmine	AD	Used in clinical treatment of Alzheimer's disease
Selegiline	AD	Used in clinical treatment of Alzheimer's disease
Valproic Acid	Epilepsy	[31]

Figures

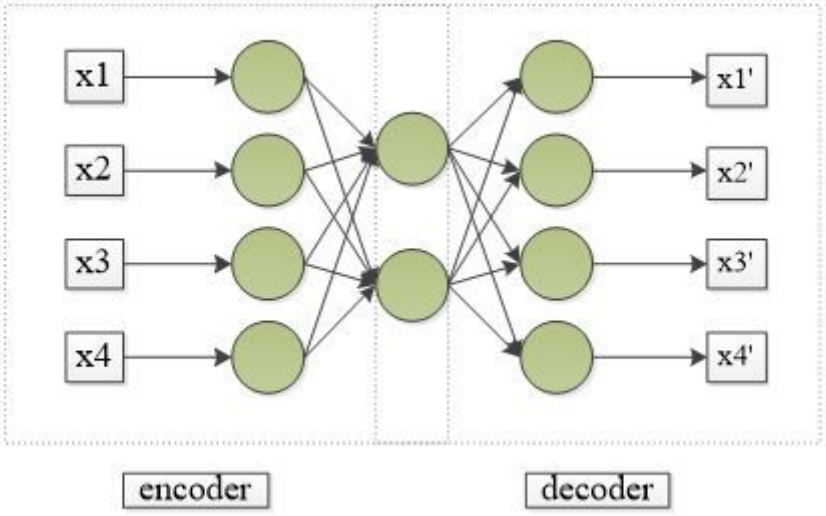


Figure 1

Structure diagram of Auto-encoder

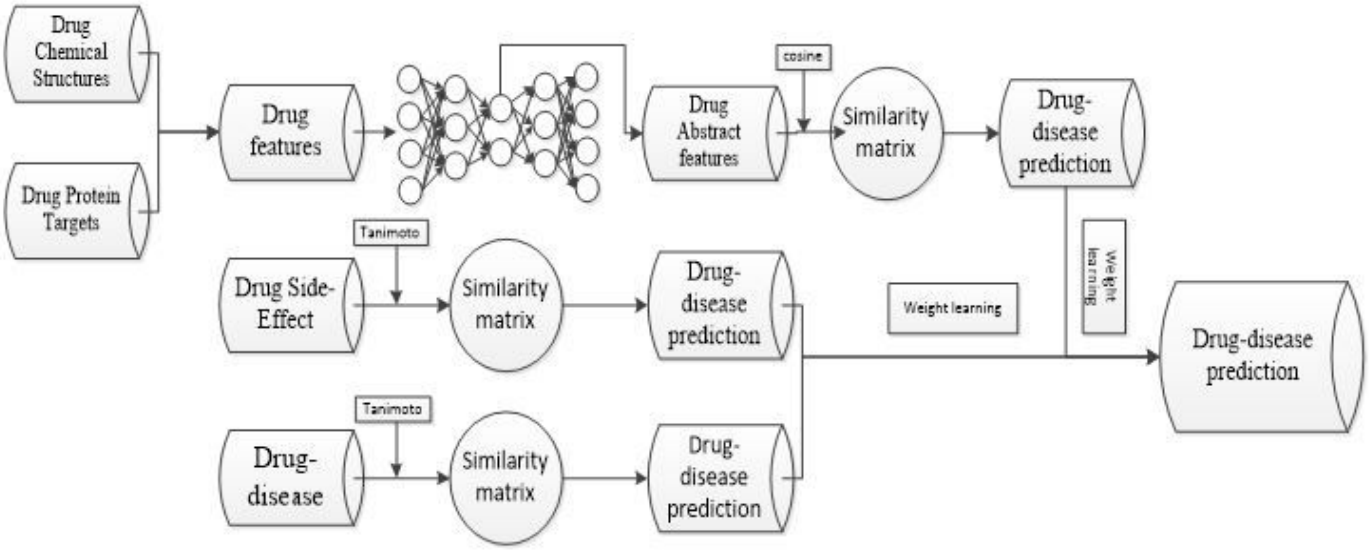


Figure 2

Framework of the drug repositioning algorithm based on deep auto-encoder and adaptive fusion. It includes four types of origin data and three types of data for experiments.

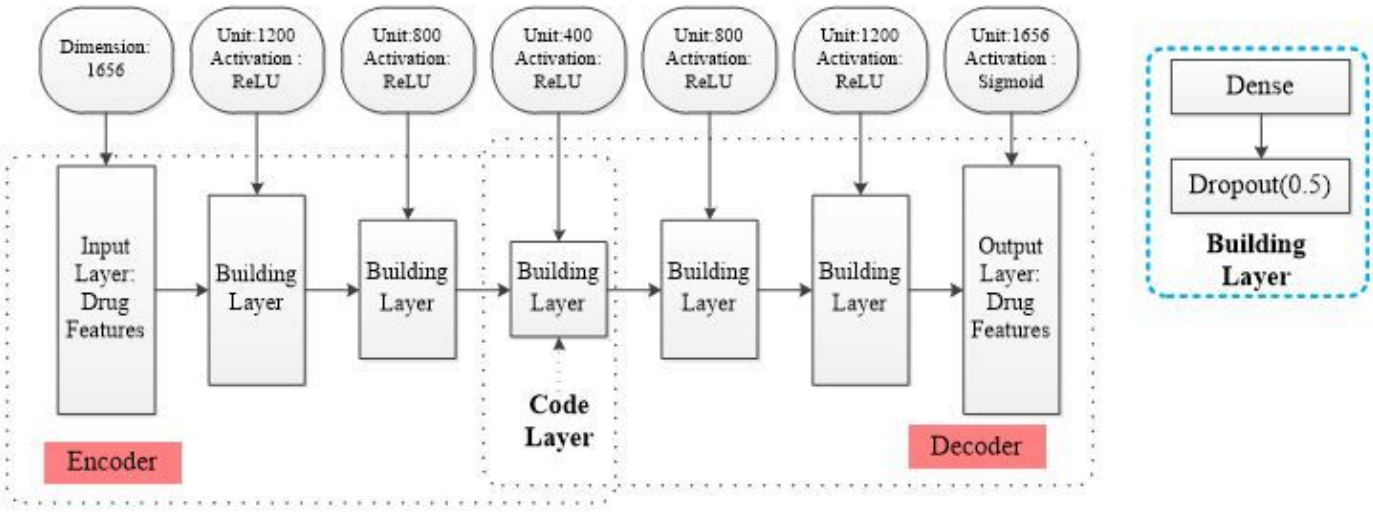


Figure 3

A schematic diagram of a deep auto-encoder framework for drug features data dimensionality reduction. There are two sections to the overall structure: one is the encoder and the other is the decoder. The encoder contains four layers: an input layer, two building layers and an encoding layer; the decoder includes four layers: an encoding layer, two building layers, and an output layer. The building layer is composed of a Dense and Dropout.

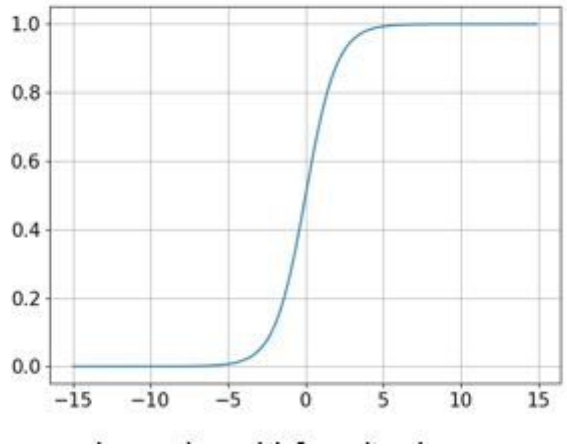


Figure 4

Sigmoid function image

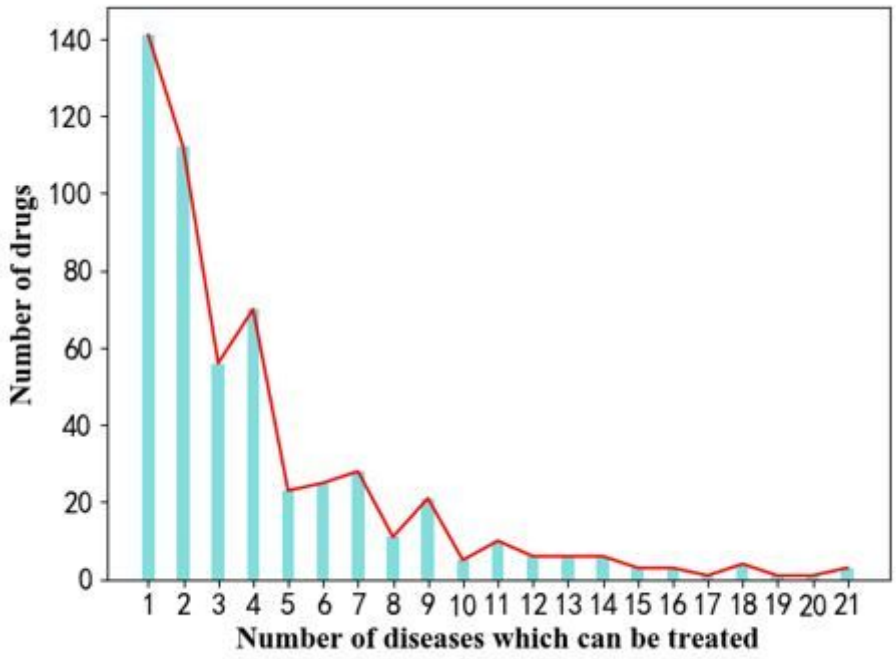


Figure 5

The number of diseases that can be treated by each drug in the drug-disease associated data

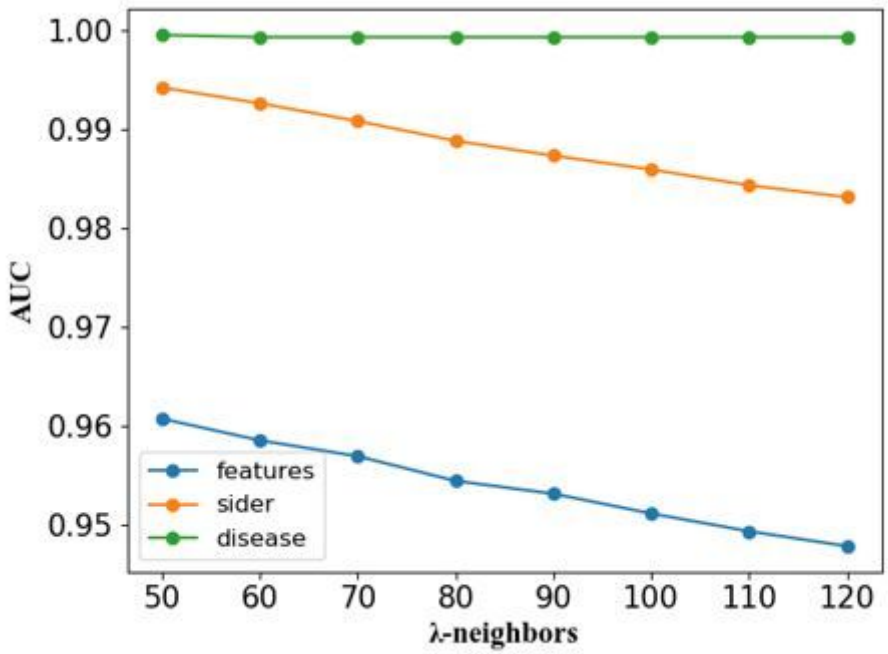


Figure 6

The figure provides the AUC value changes of the three data sources under different neighbors. In order to ensure the number of effective neighbors, we respectively give the AUC values under the number of neighbors from 50 to 120. Through comparison, it is found that the AUC value of the three data sources is the highest under the number of 50 neighbors.

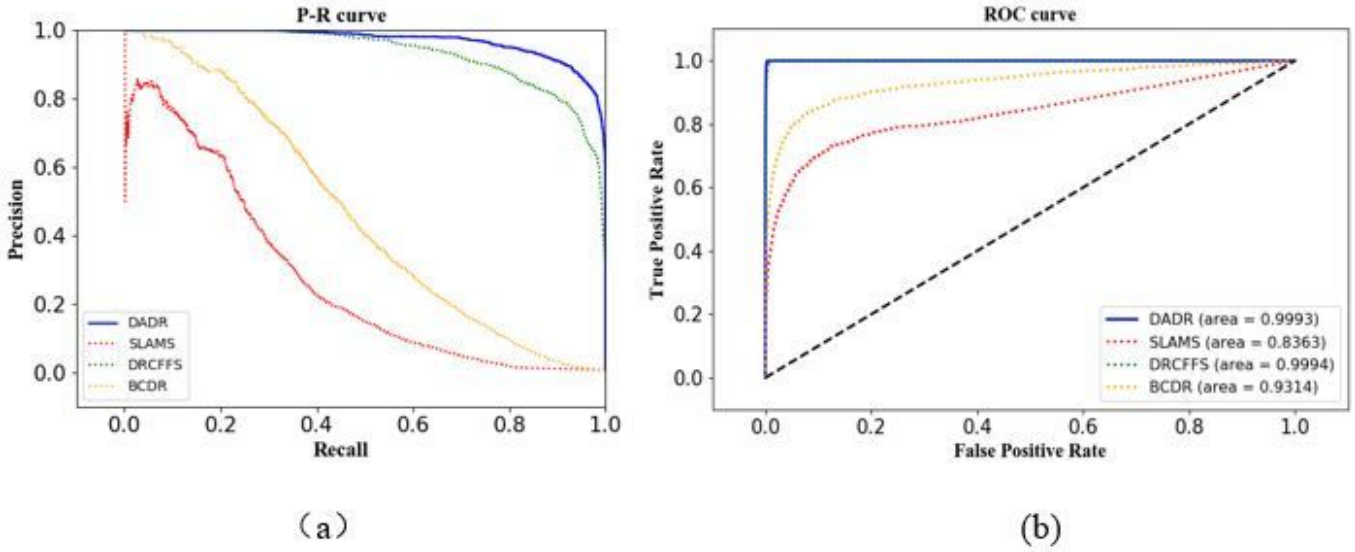


Figure 7

Figure (a) provides a the PR curves of the DRDA algorithm and the other three algorithms. The PR curve of DRDA wraps the PR curves of the other three algorithms, so the effect is the best; Figure (b) shows the ROC of the four algorithms Comparing the curve and AUC value, DADR and DRCFFS have the best effect.

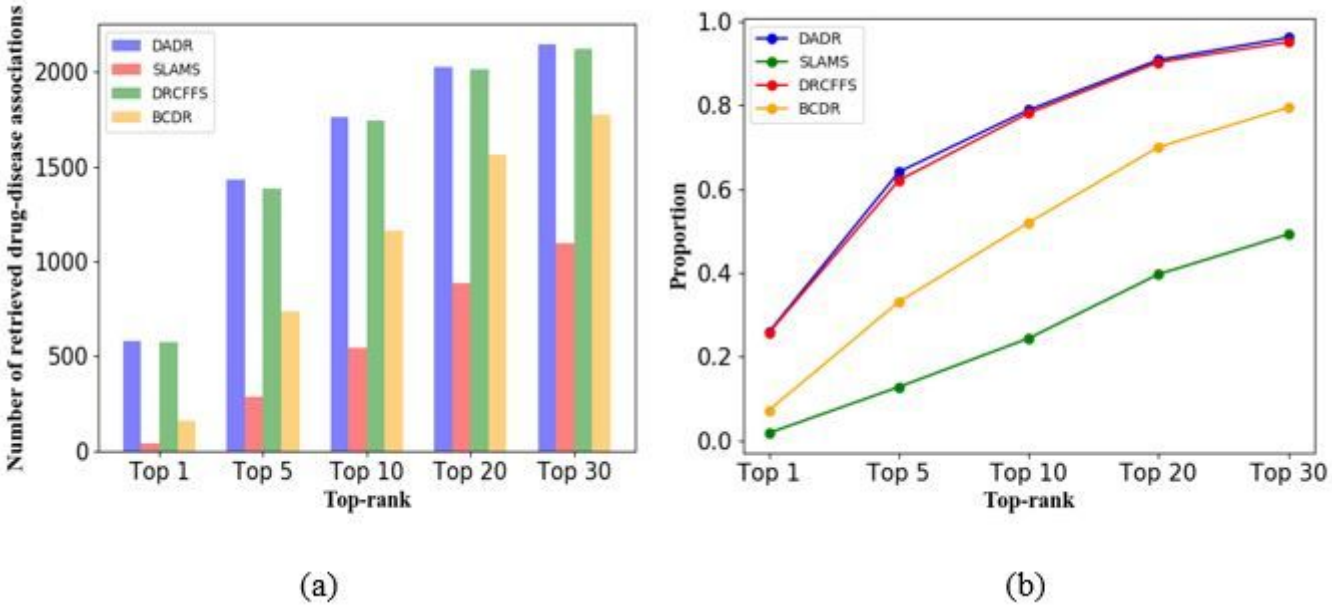


Figure 8

Figure (a) provides correct prediction of the drug-disease associated number with known treatment relationships for the four algorithms at different thresholds. Figure (b) provides the proportion of the number of drug-disease associations correctly predicted by the four algorithms with known treatment relationships at different thresholds to all drug-disease with known treatment relationships