Deep collaborative filtering for prediction of disease genes

Xiangxiang Zeng, *Senior Member, IEEE,* Yinglai Lin, Yuying He, Linyuan Lü, Xiaoping Min*, and Alfonso Rodríguez-Patón

Abstract—Accurate prioritization of potential disease genes is a fundamental challenge in biomedical research. Various algorithms have been developed to solve such problems. Inductive Matrix Completion (IMC) is one of the most reliable models for its well-established framework and its superior performance in predicting gene-disease associations. However, the IMC method does not hierarchically extract deep features, which might limit the quality of recovery. In this case, the architecture of deep learning, which obtains high-level representations and handles noises and outliers presented in large-scale biological datasets, is introduced into the side information of genes in our Deep Collaborative Filtering (DCF) model. Further, for lack of negative examples, we also exploit Positive-Unlabeled (PU) learning formulation to low-rank matrix completion. Our approach achieves substantially improved performance over other state-of-the-art methods on diseases from the Online Mendelian Inheritance in Man (OMIM) database. Our approach is 10% more efficient than standard IMC in detecting a true association, and significantly outperforms other alternatives in terms of the precision-recall metric at the top-k predictions. Moreover, we also validate the disease with no previously known gene associations and newly reported OMIM associations. The experimental results show that DCF is still satisfactory for ranking novel disease phenotypes as well as mining unexplored relationships. The source code and the data are available at https://github.com/xzenglab/Deep-Collaborative-Filtering.

Index Terms—Disease genes prediction, Data integration, Deep learning, PU learning, Matrix completion.

1 Introduction

IDENTIFYING gene-disease links are of great value in human disease diagnosis and treatment. The known disease-associated genes reported to public databases, such as the Online Mendelian Inheritance in Man (OMIM) and the Genetic Association Database [1], represent a small fraction of actual relationships. Hence, discovering disease genes remains important. Traditional gene-mapping approaches involve linkage analysis and a genome-wide association study (GWAS) [2]. Due to the limited number of crossovers in sampled families, linkage analyses usually identify only chromosomal intervals that may contain up to hundreds of candidate genes. Genome-wide association studies may also return many regions which remain to be examined in future works. Experimental validations of so many candidate genes are time-consuming and expensive.

Because integrating multiple auxiliary sources of data is essential to gene-disease traits, a series of network-based computational alternatives have been proposed in the past decade [3], [4], [5], [6]. The common motivation of these methods is that genes causing the same or similar diseases tend to closely correlate with one another in the biological networks. The typical evidence that these models can be categorized is as follows [7]: text-mining of biomedical literature, functional annotations, pathways and ontologies, phenotype relationships, intrinsic gene properties, sequence data, protein-protein interactions, regulatory information, orthologous relationships, and gene expression information. For example, [8] used a text-mining approach to classify large-scale human phenotypes contained in the OMIM database. To define the similarities in protein-protein interaction networks, [9] inferred gene-disease connections by using a global network distance measure called random walk analysis. Besides, more recently [6] conducted CATAPULT and Katz on the heterogeneous network, which develops an extension of the random walk using walk count to compute similarities between two objects and incorporate relevant information from other species, such as fruit fly and mouse. The main drawback of those network-based methods is that they fall short of generalizing novel diseases, for which there are no gene linkage studies yet. In this regard, [10] constructed a resulting method, Inductive Matrix Completion (IMC), based on different biological sources, which can be applied to diseases not seen at training time. Of the above methods for prioritizing genes relevant to a given disease, standard IMC performs the best, even though it may result in a shallow understanding of the features.

The advancement of deep learning leverages is widely applied to classification tasks and representations learning. These deep systems with multiple layers have been shown to yield promising performance in extracting more strong features of data such as [11], [12]. These projects directly use convolutional neural network (CNN) [11], or deep belief network (DBN) [12], to obtain latent factors

- *Xiangxiang Zeng, Yinglai Lin, Yuying He and Xiaoping Min (corresponding author) are with Department of Computer Science, Xiamen University, Xiamen, China, 361005. (Xiangxiang Zeng, Yinglai Lin, Yuying He are co-first authors.)*
- *Linyuan Lü is with Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, and Alibaba Research Center for Complexity Sciences, Alibaba Business College, Hangzhou Normal University, Hangzhou 311121, China.*
- *Xiangxiang Zeng and Alfonso Rodríguez-Patón are with Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Campus de Montegancedo, Boadilla del Monte 28660, Madrid, Spain.*

received date:??

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2907536, IEEE/ACM Transactions on Computational Biology and Bioinformatics

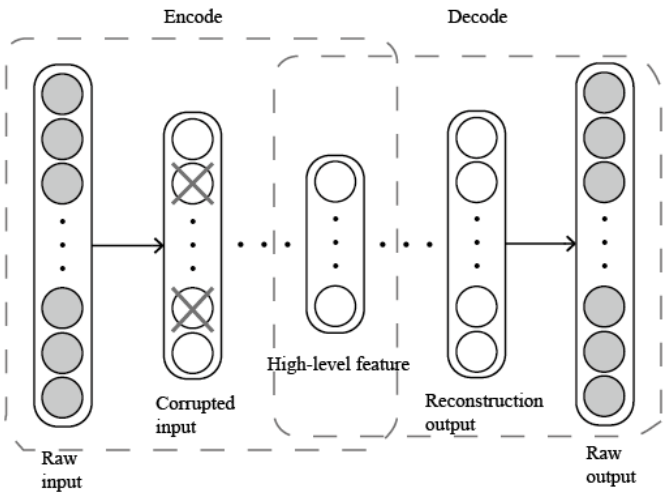


Fig. 1: Illustration of Stacked Denoising Autoencoders.

for content information. Refs. [13] and [14] use Bayesian stacked denoising autoencoders (SDAE), and marginalized SDAE, for collaborative filtering. In this paper, we propose a flexible and robust model, which jointly uses deep latent factors of genes and disease learning from side information and perform collaborative filtering from the gene-disease associations matrix. Further, an effort is made to modify the optimization. For any given phenotype, it is difficult to verify that the unobserved evidence indicates whether such a connection is nonexistent or hidden. Inspired by the various learning methods focused on PU learning for matrix completion in previous literature [15], [16], we employed PU learning in our work. Experiments show that our hybrid model considerably outperforms the state of the art. Specifically, the main contributions of this paper are as follows:

We summarize our major contributions as follows. First, we propose a Deep Collaborative Filtering (DCF) model, which integrates deep representation learning and matrix completion. Second, we present Stacked Denoising Autoencoders (SDAE) as the deep learning algorithm to reduce dimensionality to a manageable size of genes with relevant information. Besides, SDAE not only copes with noisy and incomplete nature from diverse data sources, but also achieves a higher compression rate that makes it more suitable for biological cases. Additionally, PU learning is adopted in processing the association matrix. Third, we compare our model with the state-of-the-art methods of OMIM diseases. Experimental results demonstrate the superiority of DCF in retrieving true relationships and precision-recall measurements. We also analyze the novelty of prediction for new diseases and new associations. By masking partial features, we explore feature importance of given gene features on our proposed framework.

This paper is organized as follows: In Section 2, a simplified review of SDAE is given first, followed by a discussion of the limitations of traditional matrix completion techniques. In Section 3, a detailed description of our proposed approach is given. Experimental datasets, evaluation metrics and results analysis are presented in Section 4. Conclusions are presented in Section 5.

2 PRELIMINARIES

Before describing our approach, we discuss the preliminaries as follows:

2.1 Stacked Denoising Autoencoders

An autoencoder (AE) [17] is a feedforward neural network with one hidden layer, aimed at learning compressed and distributed representations automatically in the encoding phase and reconstructing its clean input in the decoding phase. Given an input, x, the model applies the encoder, $f(\cdot)$, to map it to a hidden feature $f(x)$, while the decoder $g(\cdot)$ maps the hidden feature to the output layer $\hat{x}$ as a reconstructed version of x, such that $g(f(x)) = \hat{x}$. The parameters of the autoencoder are learned to minimize the reconstruction error, measured by some loss function $L(x, g(f(x)))$.

To force the autoencoder to generate more robust features rather than a simple identity, a denoising autoencoder (DAE) is proposed by [18]. Instead of directly using the original input, x, a DAE incorporates several different corruption criteria into raw data, including masking noise or additive isotropic Gaussian noise. The process of adding noise improves the robustness of trained features and consequently the resulting models. In this study, we adopted Stacked DAE (SDAE) proposed by [19], which stacks DAE to form a deep network, allowing SDAE to learn hierarchical features from input layer to hidden layer. A SDAE model minimizes the regularized problem and tackles reconstruction error in a manner analogous to AE, defined as follows:

$$\min_{\{W_l\},\{b_l\}} \|x - \hat{x}\|_F^2 + \lambda \sum_l \|W_l\|_F^2 \tag{1}$$

where L is the number of layers, W_l is weight matrix and b_l is bias vector of layer $l \in \{1, ..., L\}$, which can be learned by back-propagation algorithm. λ is a regularization parameter and $\|\cdot\|_F$ denotes the Frobenius norm.

The first $L/2$ layers of the model act as an encoder, and the last $L/2$ layers act as a decoder. The middle layer is the key that enables SDAE to reduce dimensionality and extract effective representations of side information.

2.2 Matrix Factorization

In this section, we first introduce low rank matrix completion and inductive matrix completion. The low rank matrix completion (MC) is one of the most popular and successful collaborative filtering methods for recommender systems [20]. The goal is to recover an underlying matrix by using the observed entities of P_{ij}, defined as follows:

$$\min_{W,H} \sum_{(i,j)\in\Omega} (P_{ij} - (WH^T)_{ij})^2 + \frac{\lambda}{2}(\|W\|_F^2 + \|H\|_F^2) \tag{2}$$

where λ is a regularization parameter and $\Omega \in N_x \times N_y$ is the observed entries from the true underlying matrix. Under the assumption of the matrix is modeled to be low rank, which means $W = \mathbb{R}^{N_x \times k}$ and $H = \mathbb{R}^{N_y \times k}$ share a low dimensional latent space, satisfying $k \ll N_x, N_y$.

2.3 Inductive Matrix Completion

Traditional matrix completion is based on the transductive setting. Also, all matrix completion approaches suffer from extreme sparsity of the observed matrix and the cold-start problem. To alleviate this limitation, an inductive matrix completion (IMC) [21] was developed, which can be interpreted as a generalization of the transductive multi-label formulation and enables us to incorporate side information. This technology was applied to make predictions on gene-disease associations [10]. In particular, side information of both entities are given in two matrices: $x_i \in \mathbb{R}^{N_g}$ denotes the feature vector for gene i and $y_j \in \mathbb{R}^{N_d}$ denotes the feature vector for disease j. The IMC assumes that the underlying association matrix is generated by applying gene and disease feature vectors to a low-rank matrix, which is learned from a training set of gene-disease associations, formulated as:

$$\min_{W,H} \sum_{(i,j)\in\Omega} \ell(P_{ij}, x_i^T W H^T y_j) + \frac{\lambda}{2}(\|W\|_F^2 + \|H\|_F^2) \quad (3)$$

where ℓ is a loss function that measures the deviation between the predictions and observations, e.g., squared loss $\ell_{sq}(a,b) = (a-b)^2$ and logistic loss $\ell_{log}(a,b) = \log(1+e^{-ab})$.

2.4 PU Learning

Existing machine learning methods typically treat the known disease genes as the positive and the unknown genes as the negative, this may lead to an over-punishment on unlabeled samples as there can be unknown disease genes in unlabeled samples. To avoid this, we treat unknown samples as unlabeled, and perform Positive-Unlabeled (PU) learning on the set. Thus, the following optimization problem is considered [15]:

$$\begin{aligned}\min_{W,H} &\sum_{i,j:P_{ij}=1} \ell(P_{ij}, WH^T) + \alpha \sum_{i,j:P_{ij}=0} \ell(P_{ij}, WH^T) \\ &+ \frac{\lambda}{2}(\|W\|_F^2 + \|W\|_F^2)\end{aligned} \quad (4)$$

where α denotes the weight of the loss function of unlabeled samples.

3 METHODS

Although the IMC is capable of impressive accuracy and integrating rich evidence for predicting gene-disease links, there is still something to be desired, especially in regard to the low-quality and high-dimensionality of biological data. Here, we present our model, Deep Collaborative Filtering (DCF). This method is based on the following recognitions: existing literature shows that SDAE is a powerful way to construct high-quality representations without requiring prior knowledge. In addition, we consider the bioinformatics data, which consists of positive gene-disease pairs (indicating that a gene is linked to a certain disease), but no available negative pairs, i.e. genes that are not related to the disease. Consequently, this problem is naturally studied in the PU learning framework, where observed and unobserved entries are penalized differently in the objective. To this end, the combination of SDAE and PU learning task that we proposed is beneficial for training more expressive models. A brief description of the DCF framework is given below.

3.1 A General Framework

With the intuition of dimensionality reduction, we embrace both SDAE and Principal Components Analysis (PCA) to obtain low-dimensional features from gene and disease side information, drawn by set X and set Y, respectively. Given a gene-disease association matrix P, DCF jointly gains compressed features from side information and learns the projection matrix $Z = WH^T$ using the observed entries from P under a PU learning setting, formulated as:

$$\mathcal{L}^+(P, XZY^T) + \alpha\mathcal{L}^-(P, XZY^T) + \frac{\lambda}{2}(\|W\|_F^2 + \|H\|_F^2) \quad (5)$$

where $\mathcal{L}^+(W,H)$ is the loss function corresponding to positive entries, and the $\mathcal{L}^-(W,H)$ is the loss of negative samples. Since standard inductive matrix completion usually ignores negative entries, to address this issue, a small value of α biases the estimation of the unobserved examples toward zero.

3.2 Deep Collaborative Filtering

Let $X = \left[x_1, ..., x_{N_g}\right]^T$ denote the trained features matrix of genes, where the i-th row is the gene i feature vector $x_i \in \mathbb{R}^{f_g}, i = 1, ..., N_g$, and let $Y = [y_1, ..., y_{N_d}]^T$ denote the trained features matrix of diseases, where the j-th row is the disease j feature vector $y_j \in \mathbb{R}^{f_d}, j = 1, ..., N_d$. In particular, SDAE and PCA are established for the equal dimensional space m on the latent features of each related dataset. For example, considering the gene functional network, $G \in \mathbb{R}^{N_g \times N_g}$, and we use $\bar{G}$ to denote the reconstructed outputs. As discussed before, the SDAE objective function works to minimize the loss between $\bar{G}$ and G to connect the clean inputs with the latent features. Now, $N_g \times m$ represents the size of low-dimensional vector representations via SDAE, whose dimensionality m is much lower than that of G. In this paper, we conduct SDAE on gene expression data, gene functional interactions and gene–phenotype associations from other species. Specifically, the term-document matrix from OMIM web pages and disease similarities retain the PCA analysis for extreme sparsity. An overview of DCF is presented in Figure 2.

The gene-disease associations matrix is given as $P \in \mathbb{R}^{N_g \times N_d}$, where N_g is the number of genes and N_d is the number of diseases. When $P_{ij} = 1$, infers gene i is linked to disease j while zero indicates the relationship is unobserved. After the feature extraction process, we construct a decomposing function to recover a low-rank matrix $Z \in \mathbb{R}^{f_g \times f_d}$ from the known associations matrix P with the form of $Z = WH^T$, where $W \in \mathbb{R}^{f_g \times k}$ and $H \in \mathbb{R}^{f_d \times k}$, $k \ll N_g, N_d$. The optimization problem our model is parameterized as:

$$\begin{aligned}\min_{W,H} &\sum_{(i,j)\ \in\Omega^+} (P_{ij} - x_i W H^T y_j^T)^2 \\ &+\alpha \sum_{(i,j)\ \in\Omega^-} (P_{ij} - x_i W H^T y_j^T)^2 + \frac{\lambda}{2}(\|W\|_F^2 + \|H\|_F^2)\end{aligned} \quad (6)$$

where the set Ω includes both positive and negative entries, such that $\Omega = \Omega^+ \bigcup \Omega^-$, let Ω^+ denote the observed samples and Ω^- denote the missing entries chosen as negatives. For biased inductive matrix completion, the value α is the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2907536, IEEE/ACM Transactions on Computational Biology and Bioinformatics

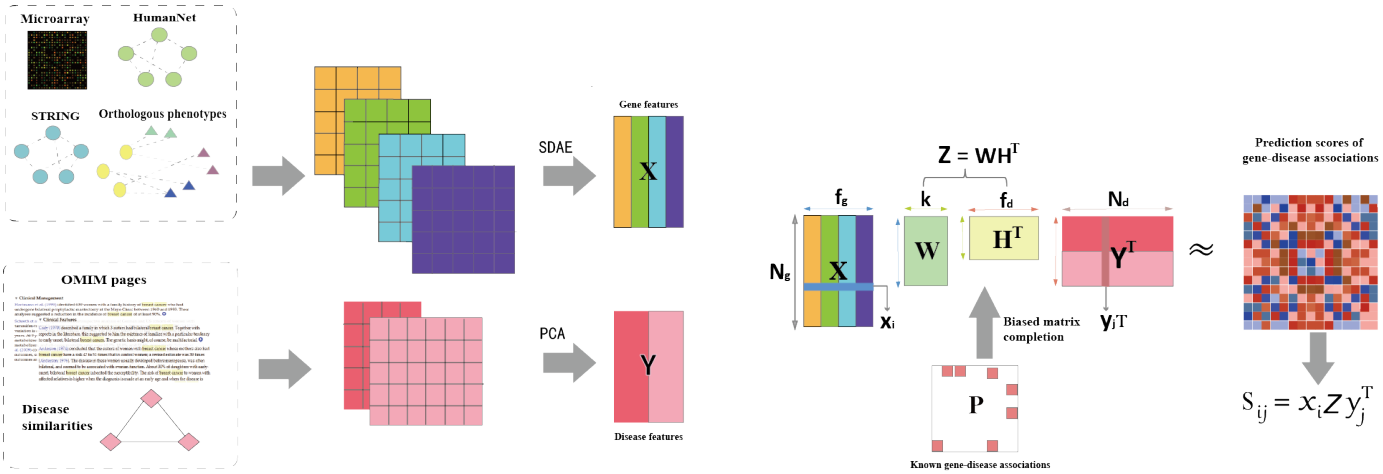


Fig. 2: The flowchart of the proposed approach. DCF first maps a variety of gene and disease-related sources to a low-dimensional space and constructs informative representations using dimensionality reduction algorithm SDAE and PCA. Then applies a biased scheme PU learning to perform matrix completion on known gene-disease associations.

key parameter so that $\alpha < 1$ because the penalty weights for observed entries must be greater than the missing ones. In our experiment, the biased value α and regulation parameter λ are selected over the grid search.

Then, we approximate the likelihood of the pairwise interaction score between gene i and disease j as:

$$score(i, j) = x_i W H^T y_j^T \tag{7}$$

where the higher score suggests a higher chance that gene i is correlated with disease j.

3.3 Optimization

The optimization of the objective function becomes convex if we alter one variable and fix the others. To do so, we fix W and update H (and vice versa) to optimize (6). Two major approaches to choose negative entries are: (1) Subsample a set with a size similar to that of the observed positive entries ("Subsampled" method) or (2) treat all missing entries as negative ("Full" method). Generally, the Full approach performs better than the Subsampled approach, but often inefficient due to a large number of negative samples. Recently, [22] developed an efficient optimization method for the Full method, by sequentially optimizing a pair of columns from W and H rather than optimizing all the values in W and H. Thus, we adopt their method and extend it to PU learning. To approximately solve (6), we exploit conjugate gradient descent with the squared loss in the iterations.

4 Results and Discussions

In this section, we evaluate the performance of our DCF model with five real-world datasets in biology and compare our DCF with the state-of-the-art algorithms.

4.1 Dataset

- **auxiliary data for gene feature extraction**
  - **Microarray** measurements of gene expression levels in different tissue samples, obtained from BioGPS (www.biogps.org) and Connectivity Map (www.broadinstitute.org/cmap) by [10]. The detail for microarray and other gene/disease is shown in Table1.
  - **HumanNet** is a large-scale functional gene network which incorporates diverse data sets, including mRNA expression, protein-protein interactions, protein complex data, and comparative genomics (but not disease or phenotype data) [23].
  - **Gene-phenotype associations** from other species that are relatively richer compared to gene-disease studies in humans. We use phenotype for eight different species: plant [24], worm [25], [26], fruit fly [27], mouse [28], yeast [29], [30], [31], [32], Escherichia coli [33], zebrafish [34] and chicken [35].The details is summarized in Table 2.
  - **STRING** [36] is a gene-gene interaction networks with large size, downloaded from NDEx (http://www.ndexbio.org), processed by [37], which quantitatively incorporates several studies and interaction types and adjusting the weight for each gene pair on the total weight of evidence.
- **auxiliary data for disease feature extraction**
  - **Disease similarities network** MimMiner [8] gives a weighted network between diseases.
  - **The web pages for the OMIM diseases** used in [10], including the 'Clinical Features' and 'Clinical Management' sections of the web pages that document the symptoms, medication and responses by patients, and related studies of effects of different courses of therapies.
- **known gene-disease links**
  - **Gene-disease association** is contained as evaluation data, downloaded from the Online Mendelian Inheritance in Man (OMIM) project on August 11, 2011 used in [6]. OMIM phenotypes have become the standard data set for evaluation of prediction of gene-disease associations [3], [4], [5], [6], [9], [10], [38], [39].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2907536, IEEE/ACM Transactions on Computational Biology and Bioinformatics

index	gene/disease name	# gene/disease
1	Microarray	8,755
2	HumanNet	18,714
3	STRING	18,266
4	Disease similarities	3,209
5	OMIM	3,209

TABLE 1: Detail for gene/disease. The table shows size of gene/disease used in our experiments.The size of Disease similarities and OMIM is far larger than 3209, but we only care about some diseases.

index	Species name	#Phenotypes	#Associations
1	Human	3,209	3,954
2	Plant	1,137	12,010
3	Worm	744	30,519
4	Fruit Fly	2,503	68,525
5	Zebrafish	1,143	4,500
6	Escherichia coli	324	72,846
7	Chicken	1,188	22,150
8	Mouse	4,662	75,199
9	Yeast	1,243	73,284

TABLE 2: Detail for species. The table shows size of species used in our experiments.

4.2 Evaluation metric

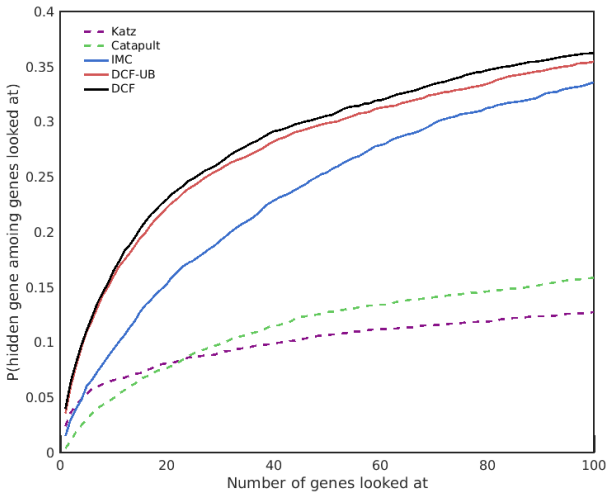
Similar to ProDiGe [38], CATAPULT [6] and IMC [10], as a measure to compare the performance of different methods, we employ the cumulative distribution function (CDF) of the ranks (i.e., the probability that the rank at which a hidden gene-disease pair is retrieved is less than a threshold r) . Here, we report results for $r \leq 100$.

We also evaluate and compare the different methods using the more familiar precision and recall measures. Precision measures the fraction of true positives (genes) recovered in the top-r predictions. Recall is the ratio of true positives recovered in the top-r predictions for a trait to the total number of true positives in the hidden set. Presented in Figure 3(b) is the plot of precision-recall rates for different values of thresholds, r , in the range $0 \leq r \leq 100$.

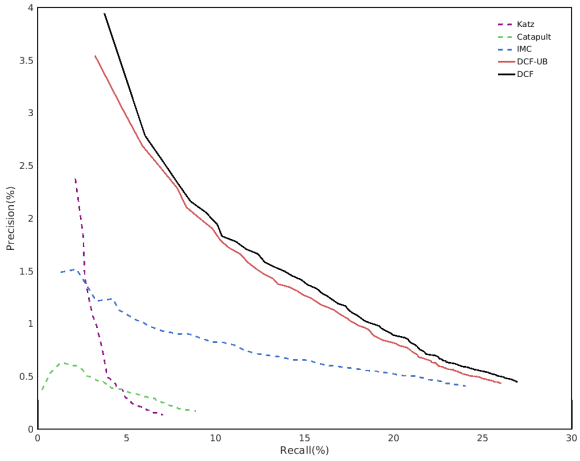
4.3 Baselines and Experimental Settings

The prediction results are based on detailed comparisons between the models listed as follows:

1) *Katz:* The Katz measure [6] is a graph-based method for finding nodes similarity to a given node by computing how many paths of different lengths exist between the pair.
2) *CATAPULT:* Combing dATa Across species using Positive-Unlabeled Learning Techniques [6] is a supervised learning method that trains a bagging support vector machine classifier over a heterogeneous network.
3) *IMC:* Inductive Matrix Completion [10] is a novel approach developed by (Jain and Dhillon, 2013) which incorporates multiple features while trying to learn gene-disease associations.
4) *DCF-UB:* Unbiased Deep Collaborative Filtering uses SDAE to manage gene features. The process of matrix completion remains the same as that of IMC.



(a) CDF of ranks



(b) Precision-Recall curves

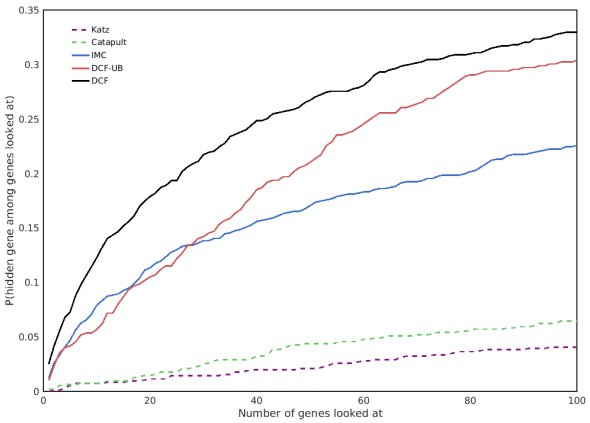
Fig. 3: Performance comparison with other state-of-the-art methods for disease gene prioritization. (a) All results are summarized over three-fold cross-validation to illustrate the probability that a true gene association is retrieved in the top-100. (b) Performance of each method is assessed by the area under precision-recall curve.

5) *DCF:* Our approach is proposed as described above. It is a hybrid model, which unifies the deep learning structure, SDAE, with the PU inductive matrix completion.

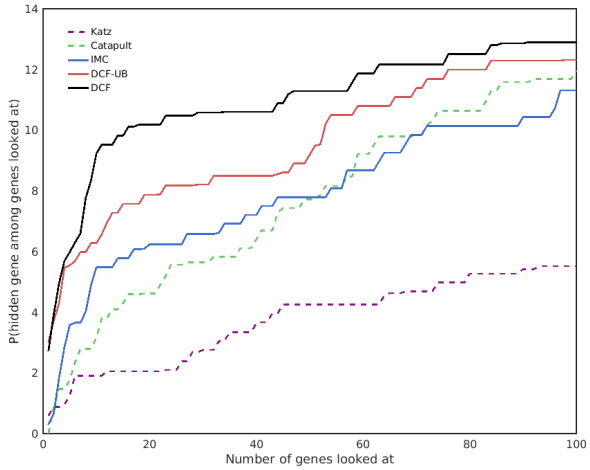
For the Microarray dataset, first we linearly transform the expression range of each gene to between zero and one, then build a 4-layer SDAE. The number of hidden units comes to '1000-500-200-100'. Masking noise is processed with a noise level of 0.1 to obtain a corrupted inputs version, in which the elements for each sample are set to zero randomly at a preset corruption rate. We adopt sigmoid for activation and dropout for generalization of every layer except the last layer, setting learning rate to be 0.1 and epoch to be 500.

For the orthologous phenotypes and gene-gene interaction network, we train deep architecture over combinations

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2907536, IEEE/ACM Transactions on Computational Biology and Bioinformatics



(a) Singleton diseases



(b) Newly discovered associations

Fig. 4: Evaluation on singleton disease and recently reported associations to show the rank of cumulative distribution and of recall, respectively. (a) Performance on singleton disease. (b) performance on 493 newly discovered associations.

of the following parameters: 2-layer network of '200-100' nodes, corruption level of 0.3, and learning rate of 0.1. To accelerate the convergence process, we train DAE in batches of samples. The number batch size is set to 128. The weight matrix and bias vector are optimized through stochastic gradient descent and updated in each batch.

Further, the dimensionality of each learned latent factor for gene and disease is set to 100; then, the rank of the low-rank matrix, k, is set to 200, accordingly. To test our method, we perform three-fold cross-validation. All gene-disease pairs are randomly partitioned into three approximately equal size sets. After a grid search, we discover that DCF performs best when the regularization parameter λ = 0.01 and α = 0.05 (α determines the penalty of the unobserved entries toward zero).

4.4 Overall Performance

As is shown in Figure 3, out of the competitive methods, our model achieves a distinct improvement in the two evaluation metrics. The three-fold cross-validation result shows that DCF has a close to 37% chance of retrieving a true gene in the top-100 predictions for disease; whereas, IMC has only about 33%. We implement CATAPULT and Katz on the combined network, which contains the gene-gene interaction network from HumanNet and gene-phenotype networks of 8 species. For CATAPULT, 16% hidden genes are ranked in the top-100; whereas, Katz retains 12%, which might be attributed to the absence of disease features.

Figure 3(b) presents the precision-recall results comparing IMC, CATAPULT, Katz and our model. DCF outperforms IMC with respect to the standard $precision@100$ and $recall@100$ measures, which demonstrates the strength of the latent vectors learned by deep architecture and validates, as expected, that SDAE plays a key role in improving predictive power. To the best of our knowledge, gene expression data is highly noisy and correlated. Compared with the frequently used feature compression method, PCA, results reflect that SDAE deals better with the computational challenges arising from the incomplete and noisy biological data.

In addition, DCF-UB is a deep method without PU learning in its collaborative filtering phase. Figure 3 shows that the biased strategy, used in the DCF model, works to obtain better results in the gene-disease identification task.

4.5 Induction on new disease

We also underline the power of discovering new disease from the proposed DCF method. A disease with only one known association in the dataset, but none in the training (a singleton disease), may shed light on the capacity to extract effective features from side information. Figure 4(a) gives the performance comparison for this task. The CDF of ranks for the DCF illustrates superiority to predict the top-100 genes associated with novel diseases.

4.6 Evaluation on newly discovered associations

Because cross-validation on retrospective data probably leads to overoptimistic results, cross-validation is inappropriate for this issue. To evaluate the ability of the models to predict newly discovered associations, we train all methods using 12,331 genes linked to 3,209 OMIM diseases collected until May 2007 [40] and execute testing on new emerging associations from May 2007 to July 2014. The new edition between May 2007 and July 2014, carries 493 newly discovered associations with 242 genes that were not associated before. Thus, this experiment also evaluates the performance of recommending novel genes. The performance of all the methods in this validation is displayed in Figure 4(b). We can see that DCF achieves the best performance and DCF-UB performs better than IMC. Moreover, compared with the results in Figure 3(a), obvious decreases in performance of all methods are observed, due to the fewer associations known on training sets.

4.7 Investigation on feature importance for gene

In this subsection, we study the importance of 4 types of gene features, by randomly replacing some features with the full-zero matrix with the same size. As we can see in Figure

5, the STRING is the most important signal for disease gene prediction, and the orthologous is the next. Furthermore, more gene features improve the performance.

Consequently, by seamlessly combining the SDAE model for auxiliary side information and the collaborative filter for the gene-disease associations matrix, our model learns a much more meaningful representation for each gene and disease and provide more accurate prediction.

5 Conclusion

In this paper, we presented a deep collaborative filtering model that bridges SDAE and matrix completion. This method also offers a scalable platform for integrating various categories of features of both genes and diseases. Quantitative results reveal our method, by taking advantage of deep architectures, outperforms other state-of-the-art baselines. Regarding further work, the squared loss used in the DCF is unstable with the noise of the dataset. The score matrix entries equal zero or one, so there might be a more suitable loss function for the binary matrix.

In this study, we simply combine features obtained independently from microarray, gene associations networks and gene-phenotype network. However, this combination may ignore the intrinsic relationship among features that are shared with different auxiliary data. For future works, deep canonical correlation analysis [41] may be a promising method to extract features that are shared by different types of auxiliary data (views). On the other hand, other deep learning models, such as convolutional neural networks, could be unified in our DCF framework to further boost performance.

Acknowledgments

The work was supported by the National Natural Science Foundation of China (Grant Nos. 61472333, 61772441, 61872309, 61472335, 61272152, and 41476118), Natural Science Foundation of the Higher Education Institutions of Fujian Province (No. JZ160400), President Fund of Xiamen University (No. 20720170054). X. Zeng is supported by Juan de la Cierva position (code: IJCI-2015-26991). A. Rodríguez-Patón is partially supported by Spanish TIN2016-81079-R, (MINECO AEI/FEDER, EU) and Madrid Gov. B2017/BMD-3691, InGEMICS-CM(FSE/FEDER, EU).

References

[1] K. G. Becker, K. C. Barnes, T. J. Bright, and S. A. Wang, "The genetic association database." *Nature Genetics*, vol. 36, no. 5, pp. 431–432, 2004.

[2] D. Botstein and N. Risch, "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease." *Nature Genetics*, vol. 33, no. 33 Suppl, pp. 228–237, 2003.

[3] X. Wu, R. Jiang, M. Q. Zhang, and S. Li, "Network-based global inference of human disease genes," *Molecular systems biology*, vol. 4, no. 1, p. 189, 2008.

[4] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating genes and protein complexes with disease via network propagation." *Plos Computational Biology*, vol. 6, no. 1, p. e1000641, 2010.

[5] Y. Li and J. C. Patra, "Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network," *Bioinformatics*, vol. 26, no. 9, pp. 1219–1224, 2010.

[6] U. M. Singh-Blom, N. Natarajan, A. Tewari, J. O. Woods, I. S. Dhillon, and E. M. Marcotte, "Prediction and validation of gene-disease associations using methods inspired by social network analyses," *PloS one*, vol. 8, no. 9, 2013.

[7] R. M. Piro and C. F. Di, "Computational approaches to disease-gene prediction: rationale, classification and successes," *Febs Journal*, vol. 279, no. 5, pp. 678–696, 2012.

[8] M. A. V. Driel, J. Bruggeman, G. Vriend, G. B. Han, and J. A. M. Leunissen, "A text-mining analysis of the human phenome," *European Journal of Human Genetics*, vol. 14, no. 5, pp. 535–542, 2006.

[9] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes," *American Journal of Human Genetics*, vol. 82, no. 4, pp. 949–958, 2008.

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

[11] S. Dieleman and B. Schrauwen, "Deep content-based music recommendation," in *International Conference on Neural Information Processing Systems*, 2013, pp. 2643–2651.

[12] X. Wang and Y. Wang, "Improving content-based and hybrid music recommendation using deep learning," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 627–636.

[13] H. Wang, N. Wang, and D.-Y. Yeung, "Collaborative deep learning for recommender systems," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 1235–1244.

[14] S. Li, J. Kawale, and Y. Fu, "Deep collaborative filtering via marginalized denoising auto-encoder," in *ACM International on Conference on Information and Knowledge Management*, 2015, pp. 811–820.

[15] C.-J. Hsieh, N. Natarajan, and I. S. Dhillon, "Pu learning for matrix completion." in *ICML*, 2015, pp. 2445–2453.

[16] N. Natarajan, N. Rao, and I. Dhillon, "Pu matrix completion with graph information," in *IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, 2016, pp. 37–40.

[17] B. Schölkopf, J. Platt, and T. Hofmann, "Greedy layer-wise training of deep networks," in *International Conference on Neural Information Processing Systems*, 2006, pp. 153–160.

[18] P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *International Conference on Machine Learning*, 2008, pp. 1096–1103.

[19] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion." *Journal of Machine Learning Research*, vol. 11, no. 12, pp. 3371–3408, 2010.

[20] C. Volinsky, C. Volinsky, and C. Volinsky, *Matrix Factorization Techniques for Recommender Systems*. IEEE Computer Society Press, 2009.

[21] P. Jain and I. S. Dhillon, "Provable inductive matrix completion," *Computer Science*, 2013.

[22] H.-F. Yu, M. Bilenko, and C.-J. Lin, "Selection of negative samples for one-class matrix factorization," in *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM, 2017, pp. 363–371.

[23] I. Lee, U. M. Blom, P. I. Wang, J. E. Shim, and E. M. Marcotte, "Prioritizing candidate disease genes by network-based boosting of genome-wide association data," *Genome Research*, vol. 21, no. 7, p. 1109, 2011.

[24] S. David, W. Christopher, L. Philippe, T. Z. Berardini, G. H. Margarita, F. Hartmut, D. Li, M. Tom, M. Robert, and P. Larry, "The arabidopsis information resource (tair): gene structure and function annotation," *Nucleic Acids Research*, vol. 36, no. Database issue, pp. D1009–D1014, 2008.

[25] N. Chen, T. W. Harris, I. Antoshechkin, C. Bastiani, T. Bieri, D. Blasiar, K. Bradnam, P. Canaran, J. Chan, and C. K. Chen, "Wormbase: a comprehensive data resource for caenorhabditis biology and genomics," *Nucleic Acids Research*, vol. 33, no. Database issue, p. D383, 2005.

[26] R. A. Green, H. L. Kao, A. Audhya, S. Arur, J. R. Mayers, H. N. Fridolfsson, M. Schulman, S. Schloissnig, S. Niessen, and K. Laband, "A high-resolution c. elegans essential gene network based on phenotypic profiling of a complex tissue," *Cell*, vol. 145, no. 3, pp. 470–482, 2011.

[27] S. Tweedie, M. Ashburner, K. Falls, P. Leyland, P. Mcquilton, S. Marygold, G. Millburn, D. Osumisutherland, A. Schroeder, and R. Seal, "Flybase: enhancing drosophila gene ontology annota-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2907536, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Fig. 5: Experimental results for evaluation of the importance of gene features. The grids in yellow indicate that the corresponding features are selected.

tions," *Nucleic Acids Research*, vol. 37, no. Database issue, p. D555, 2009.

[28] J. T. Eppig, J. A. Blake, C. J. Bult, J. A. Kadin, and J. E. Richardson, "The mouse genome database (mgd): new features facilitating a model system," *Nucleic Acids Research*, vol. 35, no. Database issue, pp. 630–7, 2007.

[29] S. S. Dwight, M. A. Harris, K. Dolinski, C. A. Ball, G. Binkley, K. R. Christie, D. G. Fisk, L. Issel-Tarver, M. Schroeder, and G. Sherlock, "Saccharomyces genome database (sgd) provides secondary gene annotation using the gene ontology (go)," *Nucleic Acids Research*, vol. 30, no. 1, pp. 69–72, 2002.

[30] T. L. Saito, M. Ohtani, H. Sawai, F. Sano, A. Saka, D. Watanabe, M. Yukawa, Y. Ohya, and S. Morishita, "Scmd: Saccharomyces cerevisiae morphological database," *Nucleic Acids Research*, vol. 32, no. 1, pp. 319–22, 2004.

[31] K. L. Mcgary, I. Lee, and E. M. Marcotte, "Broad network-based predictability of saccharomyces cerevisiae gene loss-of-function phenotypes." *Genome Biology*, vol. 8, no. 12, p. R258, 2007.

[32] M. E. Hillenmeyer, E. Fung, J. Wildenhain, S. E. Pierce, S. Hoon, W. Lee, M. Proctor, R. P. St Onge, M. Tyers, and D. Koller, "The chemical genomic portrait of yeast: uncovering a phenotype for all genes." *Science*, vol. 320, no. 5874, pp. 362–365, 2008.

[33] R. J. Nichols, S. Sen, Y. J. Choo, P. Beltrao, M. Zietek, R. Chaba, S. Lee, K. M. Kazmierczak, K. J. Lee, and A. Wong, "Phenotypic landscape of a bacterial cell," *Cell*, vol. 144, no. 1, pp. 143–156, 2011.

[34] J. Sprague, D. Clements, T. Conlin, P. Edwards, K. Frazer, K. Schaper, E. Segerdell, P. Song, B. Sprunger, and M. Westerfield, "The zebrafish information network (zfin): the zebrafish model organism database," *Nucleic Acids Research*, vol. 34, no. 1, pp. 241–243, 2003.

[35] G. W. Bell, T. A. Yatskievych, and P. B. Antin, "Geisha, a whole-mount in situ hybridization gene expression screen in chicken embryos," *Developmental Dynamics*, vol. 229, no. 3, pp. 677–687, 2010.

[36] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork *et al.*, "The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible," *Nucleic acids research*, p. gkw937, 2016.

[37] J. K. Huang, D. E. Carlin, M. K. Yu, W. Zhang, J. F. Kreisberg, P. Tamayo, and T. Ideker, "Systematic evaluation of molecular networks for discovery of disease genes," *Cell systems*, vol. 6, no. 4, pp. 484–495, 2018.

[38] F. Mordelet and J. P. Vert, "Prodige: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples," *Bmc Bioinformatics*, vol. 12, no. 1, pp. 389–389, 2011.

[39] S. Karni, H. Soreq, and R. Sharan, "A network-based method for predicting disease-causing genes," *Journal of Computational Biology*, vol. 16, no. 2, pp. 181–189, 2009.

[40] M. Xie, T. Hwang, and R. Kuang, *Prioritizing Disease Genes by Bi-Random Walk*. Springer Berlin Heidelberg, 2015.

[41] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International Conference on Machine Learning*, 2013, pp. 1247–1255.