Hindawi
BioMed Research International
Volume 2024, Article ID 9856095, 1 page
https://doi.org/10.1155/2024/9856095

Retraction

Retracted: Heterogeneous Network-Based Inductive Matrix Methods for Predicting Biomedical Gene Disease

BioMed Research International

Received 8 January 2024; Accepted 8 January 2024; Published 9 January 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] P. Das, L. Kumar, S. Degadwala, M. N. Alam, V. Jakhmola, and C. R. Bhat, "Heterogeneous Network-Based Inductive Matrix Methods for Predicting Biomedical Gene Disease," *BioMed Research International*, vol. 2023, Article ID 7121514, 13 pages, 2023.

Hindawi
BioMed Research International
Volume 2023, Article ID 7121514, 13 pages
https://doi.org/10.1155/2023/7121514

Research Article

Heterogeneous Network-Based Inductive Matrix Methods for Predicting Biomedical Gene Disease

Pranjit Das,[1] **Loveleen Kumar**,[2] **Sheshang Degadwala**,[3] **Md. Nasre Alam**,[4] **Vikash Jakhmola**,[5] **and C. Rohith Bhat**[6]

[1] *Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (K L University), Vaddeswaram, India*
[2] *Department of Computer Science & Engineering, Swami Keshvanand Institute of Technology, Management & Gramothan, Jaipur, Rajasthan, India*
[3] *Department of Computer Engineering, Sigma Institute of Engineering, Vadodara, Gujarat, India*
[4] *Woldia University, Woldia, Ethiopia*
[5] *Uttaranchal Institute Pharmaceutical Sciences, Uttaranchal University, Dehradun, India*
[6] *Department of Computer Science and Engineering, Saveetha School of Engineering (SIMATS), Chennai, Tamilnadu, India*

Correspondence should be addressed to Md. Nasre Alam; nasarhi@wldu.edu.et

Received 22 July 2022; Revised 19 November 2022; Accepted 7 April 2023; Published 24 April 2023

Academic Editor: Abolfazl Mehbodniya

Prediction of gene-disease associations has grown in popularity in recent biomedical research. However, positive and unlabeled (PU) issues and limited gene-disease association data are common concerns with present association prediction algorithms. A gene-disease association prediction approach based on Katz-enhanced inductive matrix completion is suggested in light of the abovementioned flaws. Preestimate based on the Katz technique and refined estimation based on the inductive matrix completion approach makes the model. The Katz technique is utilized to preestimate the gene-disease association on the basis of gene-disease heterogeneous network to mitigate the effects of association data-sparse and PU issues. The Katz technique, however, necessarily introduces some noise when predicting gene-disease connections due to the similarity network's quality limitations. Therefore, the elastic net regularization approach is utilized to increase the resilience of the conventional inductive matrix completion model. As a result, the prediction effect of gene-disease connections is increased using robustness and a better inductive matrix completion model. The experimental findings demonstrate that the proposed model has dramatically increased recall and precision compared to widely used gene-disease association prediction approaches. It can also resolve the typical cold-start issue in association prediction. The proposed KIMC method may consider integrating more diverse biological data sources in the future and also aid in the effective extraction of the feature data of genes and diseases with higher correlation from this biological data to improve the prediction effect.

1. Introduction

Diseases are related to many factors, such as heredity and living environment, and many diseases are closely associated with specific genes. For example, common cancers in life [1–3], Alzheimer's disease [4], and diabetes [5], are all infections caused by a variety of gene defects. Therefore, the exploration of disease-causing genes is very crucial in understanding the causes of diseases, clinical diagnosis of conditions, and early preventive treatment. It is also a key objective for human genome research and has major implications for science and society. Moreover, the initial identification of disease-related pathogenic genes is crucial for the development of disease treatment strategies and medications.

Early gene-disease association studies were carried out based on clinical and biological experimental methods usually that consume a lot of workforces and material resources. This limits the potential researches on pathogenic gene and seriously affects the related public datasets—data quality.

For example, genetic association databases [6] and the widely used OMIM [7] dataset both record only a small number of established gene-disease associations. It is not known whether there is an association relationship between most genes and diseases, which on the one hand, leads to highly sparse known association data between genes and conditions in the dataset and, on the other hand, leads to severe data skew problems in the dataset. That is, these datasets only contain certain gene-disease relationships (referred to as positive relationships in this paper) and do not contain any determining nonassociated relationships between genes and diseases (referred to as negative relationships in this paper). For those unknown gene-disease associations (referred to in this paper as unlabeled relationships), it is necessary to predict whether there is an association between them. This kind of problem is usually called a positive and unlabeled (PU) learning problem in machine learning. Existing research has shown that the lack of negative relationships will seriously affect the learning effect of PU learning problems [8].

In recent years, through high-throughput sequencing, biomedical text mining, and other means, valuable biological information (i.e., intrinsic gene characteristics, intergene similarity information, gene array information, and disease similarity information) can be obtained. The emergence of such details also allows one to study new forecasting methods to alleviate the above shortcomings. Firstly, the Katz technique was developed which constructed a gene-disease heterogeneous network by integrating intergene similarity information, interdisease similarity information, and gene-disease association information prediction to alleviate the drawback of data sparsity. However, this method cannot effectively predict nodes that are not connected to the network and will be affected by the quality of the constructed network [9, 10]. Literature [11] turned to the popular inductive matrix completion (IMC) method in machine learning to predict gene-disease associations, effectively overcoming the cold start problem. However, this method suffers from data sparsity and the PU problem. In view of the aforementioned shortcomings, a gene-disease association prediction method based on Katz-enhanced inductive matrix completion is recommended. The model is created by preestimating using the Katz technique and fine-tuning estimation using the inductive matrix completion approach. In order to lessen the effects of association data-scarcity and PU problems, the Katz technique is used to preestimate the gene-disease association based on the gene-disease heterogeneous network.

In response to the above problems, this paper proposes a Katz-boosted inductive matrix completion for gene-disease association prediction (KIMC) model based on Katz-enabled inductive matrix completion. The motivation of this model is to use the traditional Katz method to optimize the newly proposed inductive matrix completion method, which is essentially a step-by-step gene-disease prediction paradigm, including Katz method-based preestimation and inductive matrix completion method. The refinement of the estimate mainly consists of two steps. Specifically, the Katz method was first used to predict the association of unlabeled relationships for all gene-disease pairs based on the constructed gene-disease heterogeneous network. Since data is close to 1 in the estimated association score, data can be regarded as positive association information, and data comparable to 0 can be regarded as negative association information. Katz's preestimation not only alleviates the data sparsity defect but also alleviates the PU problem implications for subsequent inductive matrix completion methods. However, limited by the quality of the constructed gene-disease heterogeneous network, the predicted gene-disease association information based on the Katz method inevitably contains a certain degree of noise. To overcome the influence of these noises on the inductive matrix completion method, this paper introduces the elastic net regularization [12] into the newly proposed inductive matrix completion method to enhance its robustness, then uses the improved elastic net regularization to induce the type matrix completion model to refine gene-disease association prediction effects. Experiments on the OMIM dataset show that the KIMC method proposed in this paper not only significantly improves recall and precision compared with several other competing approaches but also solves the typical cold start in gene-disease association prediction.

The main contributions of this paper are as follows:

(1) A gene-disease association prediction model based on Katz-enhanced inductive matrix completion is proposed. The model not only combines the advantages of the Katz method and the inductive matrix completion method but also enhances the noise-tolerant performance of the model by introducing an elastic net regularization mechanism, which can effectively alleviate the data sparsity and PU problems that traditional methods are susceptible to

(2) An efficient elastic net regularization inductive matrix completion optimization algorithm is designed using the nearest neighbor forward-backward splitting technique, and the algorithm's convergence is proved theoretically

(3) Multiple sets of experimental results on the OMIM dataset show that the proposed KIMC model can achieve better prediction results than existing prediction methods and solve the cold-start problem of effectively predicting new diseases or new genes

2. Related Works

Many disease-causing gene prediction algorithms based on different gene-disease datasets have been proposed in the past decade. These algorithms are mainly divided into methods based on network similarity measurement and techniques based on machine learning.

Literature [13] proposed the correlating protein interaction network and phenotype network to predict disease genes (CIPHER) method which hypothesized that two genes closer to the interaction network might lead to more similar diseases. Disease similarity can be explained in terms of genetic similarity, using the entire disease similarity network and PPI (protein-protein interaction) network calculation to get a score; this score measures how likely a gene is to cause a particular disease. Literature [14] improved the random walk method and proposed a random walk with restart on the heterogeneous network (RWRH) model. In this model, the gene-disease heterogeneous network is constructed using intergene similarity information, interdisease similarity information, and gene-disease association information. This method fully considers the global knowledge of the whole network. A random walk particle is used to diffuse along network connections to capture the similarity between nodes to calculate the relationship between genes and diseases. Literature [15] introduced the Katz method based on the gene-disease heterogeneous network which is widely used in social network analysis. Katz method uses the number of walk paths with different lengths between two nodes on the heterogeneous network and calculates the similarity between nodes to predict the association between genes and diseases. Literature [9] and literature [10] conducted a detailed analysis and comparison of the above methods based on network similarity measures. These methods predict genes by calculating the similarity between candidate genes and disease nodes in the network. These methods can integrate different types of gene similarity information and disease similarity information into the gene-disease heterogeneous network to enhance the amount of data information; its shortcomings are also apparent for those not connected to the heterogeneous network. Moreover, gene and disease nodes cannot be effectively predicted while relying on constructing high-quality biological network models. Based on functional gene associations and gene-phenotype connections in model organisms, two techniques for predicting gene-disease associations, the first approach, the Katz measure, is driven by its success in predicting social network links and is closely related to several of the new approaches put forth for inferring gene-disease associations. The second approach, known as CATAPULT (Combining dATa Across species using Positive-Unlabeled Learning Techniques), is a supervised machine learning approach that makes use of a biased support vector machine and features produced from walks in a heterogeneous gene-trait network. OMIM phenotypes and drug-target interactions are two different datasets that were used to evaluate the performance of the suggested methods and related state-of-the-art methodologies.

Based on the limitations of the above methods, some researchers have proposed methods based on machine learning. For example, literature [15] proposed combining data across species using positive-unlabeled learning techniques (CATAPULT) which can mine disease-causing genes by training a biased support vector machine (SVM) classifier to classify gene-phenotype associations. Since the first illness gene was discovered in 1949, thousands of other genes have been shown to be connected to various diseases [16]. The most common kind of evidence for the prediction of disease-gene connections is protein-protein interaction (PPI) networks, which have been employed in a variety of studies [17]. Prior methods attempted to predict disease-gene correlations by directly utilizing PPI networks' topological structure. However, as they only use universal PPI networks retrieved from web databases, which include a lot of false positives, the prediction accuracy cannot be increased. In order to anticipate disease-gene connections, researchers frequently integrate PPI networks with additional forms of data. Combining PPI networks with clinical data that distinguishes between patients (cases) and average people is one tactic (control) [18].

A method for predicting gene-disease associations based on Katz-enhanced inductive matrix completeness is proposed. The model is created by preestimating using the Katz technique and fine-tuning estimation using the inductive matrix completion approach. In order to lessen the effects of association data scarcity and PU problems, the Katz technique is used to preestimate the gene-disease association based on the gene-disease heterogeneous network.

microRNA-disease association (MDA) predictions have been applied since the issue was raised in the late 2000s based on the data fusion paradigm. Integrating many data sources broadens the scope of research and makes it more difficult to create algorithms that produce accurate, succinct, and consistent representations of the combined data [19]. Accurate discovery of miRNA-disease associations (MDAs), a requirement for developing successful miRNA therapies, has drawn significant scientific attention over the past 15 years, as seen by the more than 55 000 related articles that are currently available on PubMed [20]. lncRNAs have received a great deal of attention in recent years from academics all around the world. In the past several years, tens of thousands of lncRNA have been discovered in eukaryotic creatures ranging from worms to humans thanks to the rapid advancements in experimental equipment and computer prediction algorithms [21]. A family of single-stranded, covalently closed RNA molecules known as circular RNAs (circRNAs) perform a range of biological tasks. The discovery of circRNA-disease connections will aid in the diagnosis and treatment of diseases as research has shown that circRNAs are engaged in a wide range of biological processes and are crucial in the emergence of numerous complicated disorders [22].

Given a disease phenotype in question, a gene is not connected to the phenotype in question. Researchers often report positive correlations between genes and phenotypes, but negative correlations are far less common. The unlabeled gene-disease phenotype pairings function as adverse associations in the CATAPULT technique. Only the positive relationships and a significant number of unlabeled gene-disease phenotype pairings are known as negative associations, which is a peculiarity of the dataset. CATAPULT's central tenet is that the instances are not often considered evil. False negatives are severely punished, whereas false positives are not severely punished.

CATAPULT classifies the human gene-phenotype pairings with only one training session using a biased SVM. With this method, a classifier is trained to categorize the bootstrap samples as negatives alongside the positive data by selecting a random bootstrap sample of a few unlabeled examples from the set of all unlabeled examples. Positive and unlabeled samples are used by CATAPULT to create an aggregate classifier using the bagging approach.

Literature [11] proposed the IMC method, which can extract gene features from gene microarray data, gene function interaction data, and homologous gene-phenotype data of different species, from disease similarity networks and clinical manifestations of diseases. The disease characteristics are obtained from a large number of medical literatures and integrated into this method to make up for the limitation that the standard matrix completion (MC) can only rely on the existing observable associations to make predictions. It can predict new genes and diseases and solve the cold start problem encountered by the MC method. The prediction effect has been greatly improved compared to the previously proposed method. Tang et al. [23] used case studies, global and local leave-one-out cross-validation (LOOCV), and the human miRNA-disease correlation dataset derived from the HMDDv2.0 database to assess the effectiveness of DLRMC. As an outcome, the AUCs of DLRMC in global LOOCV and local LOOCV, respectively, are 0.9174 and 0.8289, which significantly beat a number of prior techniques. microRNAs (miRNAs) have been linked in numerous scientific studies to the occurrence and progression of numerous human disorders. The connection between miRNAs and human diseases has recently been the subject of an increasing amount of research. Nevertheless, the recognized connections are frequently few, and it is difficult to reliably estimate the possible associations between miRNA and diseases from vast amounts of biomedical information [24].

Identification of in silico miRNA targets is a critical step in considering that the miRNA interactome has largely not even for the most part been sufficiently mapped model creatures that were studied. There have been initiatives to promote the need for computational to support the experimental identification, and analyses are needed. This has contributed to the emergence of several miRNA target prediction methods [25], which are currently regarded as essential for the design of applicable experiments These programmes recognize in silico miRNA targets as potential research subjects in the future or for computing tasks like target enrichment analysis. Predictions made with the current computational from relevant interaction databases or web servers and algorithms can be obtained [26].

3. Preliminary Knowledge

This section mainly introduces several different gene-disease association prediction methods available. The main goal of this paper is to predict the underlying causative genes of diseases, and the gene and disease datasets used today often have only a small number of known gene-disease associations. Usually, a known gene-disease association matrix $PR^{N_g \times N_d}$ is constructed as follows:

$$P = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1 \\ & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}. \quad (1)$$

Rows and columns correspond to genes and diseases, respectively, N_g refers to the total number of genes, N_d refers to the total number of conditions, $P_{ij} = 1$ means there is an association between gene i and disease j, and $P_{ij} = 0$ denotes that the association between gene i and disease j is unknown (there may be an association or may not exist). The constructed gene-disease association matrix is highly sparse as it contains many strange associations. Moreover, there are only positive association data; therefore, the problem is referred to as a typical PU learning problem. The main task is to design effective methods to predict unknown associations to predict disease-causing genes.

3.1. Katz Method. The Katz method is similar to algorithms such as CIPHER [13] and RWRH [14], and the essence of these methods is based on the network similarity measure. Specifically, the Katz method calculates the similarity score between genes and diseases based on the gene-disease relationship network and sorts the genes corresponding to the disorders according to the similarity scores to select suitable candidate disease-causing genes. The Katz method successfully applies to social network relationship prediction [15]. It uses the number of walk paths between two nodes with different lengths to calculate the similarity between nodes. The gene and disease relationship networks are also the same. The method calculates the similarity score between nodes. Here, a gene-disease relationship heterogeneous network is constructed using the gene-gene similarity network, gene-disease association network, and disease-disease similarity network. An essential objective in bioinformatics has long been making accurate predictions of novel gene-disease correlations. The so-called guilt-by-association (GBA) approach, in which novel candidate genes are discovered through their relationship with genes previously known to be involved in the condition under study, has shown to be a particularly effective method. Direct protein-protein connections, such as those maintained by the Human Reference Protein Database (HPRD) [27], are one of the most widely used types of connection. CIPHER [28], GeneWalker [16], Prince [17] are just a few of the techniques that have been developed in recent years that have expanded the association from simply direct protein interactions to further links in various ways.

Then, the Katz method is used to predict gene-disease association in the heterogeneous network. The heterogeneous network structure is shown in Figure 1. The adjacency matrix of the illustrated heterogeneous network is expressed as

$$C = \begin{bmatrix} G & P \\ P^T & D \end{bmatrix}. \quad (2)$$

Among them, G refers to the gene-gene similarity network; D refers to the disease-disease similarity network; P refers to the gene-disease association network. Since there are not many direct associations between gene G_i and disease D_j in the network, it is necessary to express the association between genes and diseases by calculating the number of paths of different lengths between nodes.

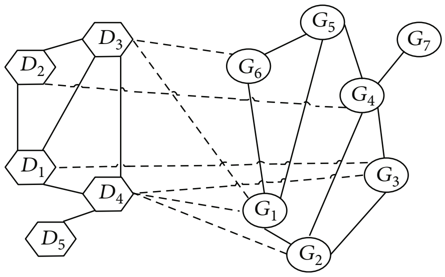


Figure 1: Structure of heterogeneous networks.

Let $(C^l)_{ij}$ represents a path of length l between gene G_i and disease D_j quantity. The internode similarity is defined on C as follows:

$$S^{\text{Katz}}(C)_{ij} = \sum_{l=1}^{k} \beta^l \left(C^l\right)_{ij}. \qquad (3)$$

Among them, β is a nonnegative constant, which is used to control the influence of paths of different lengths, and the value range of β is $(0, \min\{1, 1/\|C\|_2\})$. Converting Equation (3) into matrix form, the corresponding correlation score matrix can be expressed as follows:

$$S^{\text{Katz}}(C) = \sum_{l\geq 1} \beta^l C^l = (I - \beta C)^{-1} - I. \qquad (4)$$

However, in the Katz method, it is not necessary to consider the number of paths of all lengths because paths with shorter path lengths convey more similar information between nodes. In contrast, nodes with farther distances communicate less information, so only the sum of finite path lengths needs to be considered. An earlier study [29] found that the smaller values of k (usually $k = 3$ or $k = 4$) usually show better performance. In the experiment, taking $k = 3$ and taking out the corresponding gene-disease similarity Katz score matrix can be expressed as

$$S^{\text{Katz}}_{G\sim D} = \beta \mathrm{P} + \beta^2 (GP + PD) + \beta^3 \left(PP^T P + G^2 P + GPD + PD^2\right). \qquad (5)$$

Use Equation (5) to find the score between genes and diseases. The method integrates auxiliary information (i.e., gene-gene similarity network and disease-disease similarity network) into the gene-disease heterogeneous network, effectively improving the prediction effect. The flowchart of method implementation is presented in Figure 2.

3.2. Standard Matrix Completion (MC). Due to the apparent shortcomings of the network-based association prediction method, Katz proposed to use the matrix completion theory for gene-disease association prediction. Initially, gene-disease associations were predicted using the MC method, which decomposes the target matrix into two low-rank matrices $W \in R^{N_g \times k}$ and the product of $H \in R^{N_g \times k}$ where $k \ll N_g, N_d$. Therefore, predicting genetic disease associations can be written to solve the following optimization problem:

$$\min_{W,H} \sum_{(i,j\in\Omega)} \left(P_{ij} - W_{\mathrm{i}} H_j^T\right)^2 + \frac{\lambda}{2}\left(\|W\|_F^2 + \|H\|_F^2\right), \qquad (6)$$

where Ω is the set of positions of observed elements, λ is the regularization parameter, and W_i and H_j represent the latent features of the ith gene and the jth disease, respectively, minimizing $\lambda/2(\|W\|_F^2 + \|H\|_F^2)$, equivalent to reducing the nuclear norm of WH^T.

The gene-disease association matrix P constructed using existing biological datasets is very sparse. For instance, the OMIM database datasets show that the majority of diseases have only one gene known to be associated with them, and the majority of genes do not have any conditions that are related to them. Here, standard matrix completion cannot predict those rows and columns in the correlation matrix with no elements, i.e., suffer from the cold-start problem.

3.3. Inductive Matrix Completion (IMC). Since standard matrix completion is used to predict gene-disease associations, a single type of data (only known gene-disease associations are used), such as biomedical literature, functional annotations, protein-protein interactions, homology tables of different species, and much biometric information such as gene microarrays, cannot be effectively used. There will be a cold start problem when forecasting, and the forecasting effect is not ideal. Given the above issues, finding characteristic information that can effectively utilize such genes and diseases is necessary. The multilabel learning problem formulated in literature [30] can make good use of such feature information. In multilabel learning problem, a low-rank linear model $Z \in R^{d\times l}$ needs to be learned, in which each gene is expressed with the aid of d features and L labels. When $x \in R^d$ represents the eigenvector of the gene, the prediction for j of illness can be described as $x^T Z_j$, where Z_j represents the jth column of matrix Z.

Applying the IMC [31] model to the gene-disease association prediction problem, IMC presumes that an association matrix is constructed by using the eigenvectors w.r.t., its row, and column entities to $Z \in R^{f_g \times f_d}$ (where Z is a low-rank matrix), with the observed in P element to restore Z. Let $x_i \in R^{f_g}$, $y_j \in R^{f_d}$ denote the eigenvectors of gene i and disease j, respectively; $X \in R^{N_g \times f_g}$ refers to N_g genes, the training feature matrix, each row of which represents the eigenvector of a gene; $Y \in R^{N_d \times f_d}$ represents a feature training matrix of N_d diseases, where each row represents a feature vector for one condition. The IMC will be modelled as $P_{ij} = x_i^T Z y_i$, and the low-rank matrix Z needs to be recovered, i.e., $Z = WH^T$ where $\mathrm{W} \in R^{f_g \times k}$, $\mathrm{H} \in R^{f_g \times k}$, $\mathrm{k} \ll f_g, f_d$.

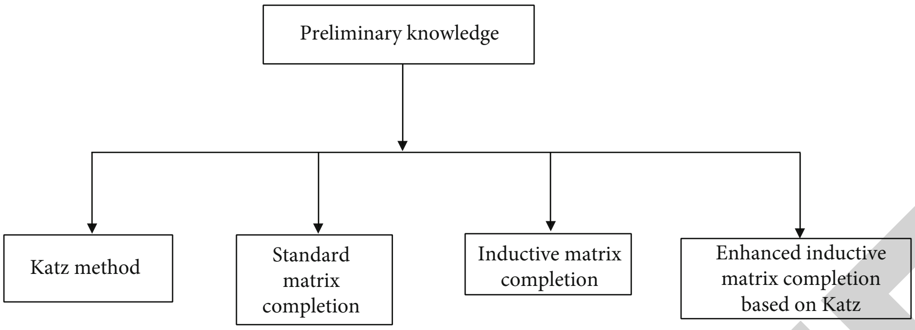

Preliminary knowledge
Katz method
Standard matrix completion
Inductive matrix completion
Enhanced inductive matrix completion based on Katz

Figure 2: Flowchart of method implementation.

Therefore, gene-disease associations predictive modelling addresses the following problems:

$$\min_{W,H} \sum_{(i,j\in\Omega)} \left(P_{ij} - x_i^T W H^T y_j\right)^2 + \frac{\lambda}{2}\left(\|W\|_F^2 + \|H\|_F^2\right). \tag{7}$$

For disease j, which does not exist in the training data, if it has its feature vector $y_j^{'}$, then for all genes i, all its associations P_{ij} can be calculated. The same is true for a new gene and can effectively solve the cold-start problem encountered by the MC method. When the count of features is large, k is set to a lower value, and the count of parameters to be learned is lesser than $f_g \times f_d$. In typical matrix completion, the number of parameters to be known is $(N_g \times N_d) \times k$. It is not difficult to discover that the parameters required for IMC learning depend simply on the number of features associated with genes and diseases rather than the number of genes and diseases.

The MC problem can be regarded as a particular case of the IMC problem when the feature matrix X of genes is a unit matrix of size N_g, and the feature matrix Y of diseases is a unit matrix of size N_d. Here, Equation (7) is solved using alternating minimization (i.e., fixed W to find H or fixed H to find W, alternating iterative solution); when any one (W or H) is selected, the answer has only one variable (H or W), and then, it can be solved by conjugate gradient descent.

3.4. Enhanced Inductive Matrix Completion Based on Katz. Due to the extreme sparseness of existing gene-disease data and the most gene-disease databases that only record identified associations, existing methods suffer from data sparsity and PU issues. Therefore, it is necessary to seek a more stable way that can alleviate the influence of the sparse problem of gene-disease association data and the power of the PU problem. Therefore, a KIMC method is proposed, which integrates the Katz method for association prediction on the gene-disease heterogeneous network and inductive matrix completion model. First, when constructing a heterogeneous network, the proven gene-gene similarity information and disease-disease similarity information can be obtained from databases widely recognized in the industry. Together with the gene-disease association information, a heterogeneous network can be formed. This type of information used by the Katz method can convey gene-disease-related information more directly than feature information. To integrate the IMC method to enhance the prediction effect without losing its inductive character, model the problem as

$$P = S_{G\sim D}^{\mathrm{Katz}}(C) + \alpha XYZ^T, \tag{8}$$

where $X \in R^{N_g \times f_g}$ represents the feature matrix of N_g genes, $Y \in R^{N_d \times f_d}$ represents the feature matrix of N_d diseases, and $Z \in R^{f_g \times f_d}$ represents the low-rank matrix that needs to be recovered. The parameter α adjusts the prediction weight using feature information, where $\alpha = 1$. Using Equation (7) to calculate $S_{G\sim D}^{\mathrm{Katz}}(C)$, the generated score data between genes and diseases, the part with a high score is regarded as positive association information, and the part with a low score is considered as negative association information.

It makes up for the shortcomings of traditional methods that can only use known gene-disease associations and can effectively alleviate the PU problem and the data sparseness problem encountered when using gene-disease association data directly. Integrate inductive matrix completion methods for residual matrix $\mathrm{R} = \mathrm{P} - S_{G\sim D}^{\mathrm{Katz}}(C)$ are solved to enhance its prediction effect. Therefore, we will use an inductive matrix complement, and the complete method solution residual R is modelled as

$$\min_{Z\in R^{f_g \times f_d}} \|Z\|_* s.t. P_\Omega\left(XYZ^T\right) = P_\Omega(R). \tag{9}$$

Due to the influence of the quality of the constructed network, the introduction of the residual matrix R will bring some noise, and the direct use of the inductive matrix completion solution will affect the prediction effect and stability. Therefore, the matrix elastic net regularization [12] is introduced to alleviate this problem. Model the solution residual R as

$$\min_{Z\in R^{f_g \times f_d}} \|Z\|_* + \frac{\lambda}{2}\|Z\|_F^2 s.t. P_\Omega\left(XYZ^T\right) = P_\Omega(R), \tag{10}$$

```
Inputs: Gene and disease feature matrices X, Y, association matrix P, set of sampling subscripts $\Omega$, gene similarity matrix G, disease similarity matrix D, parameters $\beta$, $\delta$, $\rho$, $\lambda$, and the number of iterations $\mathrm{Max}_{\mathrm{iter}}$
Output: Predicted correlation matrix $S_{G\sim D}^{Katz}(C) + XZY^T$
1. Calculate$S_{G\sim D}^{\mathrm{Katz}}$ (C) according to equation (5)
2. Calculate the residual matrix R
3. Initialize $Z_0 = 0$
4. For k= 0 to $\mathrm{Max}_{\mathrm{iter}}$
5. Update Z according to equation (17)
6. End for
7. Return $S_{G\sim D}^{Katz}$ (C) + $XZY^T$
```

Algorithm 1: Enhanced inductive matrix completion based on Katz.

Table 1: Recall vs. *r*.

Recall	Top-*r*	MC	Katz	IMC	KIMC1	KIMC2
0	0	0	0	0	0	0
0.1	20	0.05	0.1	0.2	0.15	0.25
0.2	40	0.07	0.12	0.22	0.17	0.27
0.3	60	0.06	0.11	0.21	0.16	0.26
0.4	80	0.07	0.12	0.22	0.17	0.27
0.5	100	0.08	0.13	0.23	0.18	0.28

Further, problem (10) can be transformed into an equivalent penalty function.

$$\min_{Z\in R^{f_g\times f_d}} \|Z\|_* + \frac{\lambda}{2}\|Z\|_F^2 + \frac{\rho}{2}\left\|P_\Omega\left(XYZ^T - R\right)\right\|_F^2. \tag{11}$$

This paper intends to use the nearest neighbor forward-backward splitting (PFBS) [32] technique to optimize the solution to the problem (11). May wish to order

$$\begin{aligned} F_1(Z) &= \|Z\|_*, \\ F_2(Z) &= \frac{\lambda}{2}\|Z\|_F^2 + \frac{\rho}{2}\left\|P_\Omega\left(XYZ^T - R\right)\right\|_F^2. \end{aligned} \tag{12}$$

Then, problem (11) can be formalized in the general form as follows:

$$\min_{Z\in R^{f_g\times f_d}} F_1(Z) + F_2(Z). \tag{13}$$

According to the PFBS rules, Z can be solved iteratively as follows:

$$Z^{k+1} = \arg\min_{Z\in R^{f_g\times f_d}} \left\{\delta\|Z\|_* + \frac{1}{2}\left\|Z - \left(Z^k - \delta\nabla F_2\left(Z^k\right)\right)\right\|_F^2\right\}, \tag{14}$$

where δ is the updated step size, and

$$\nabla F_2(Z) = \lambda Z + \rho\left(X^T XZY^T Y - X^T RY\right). \tag{15}$$

According to [32], for a matrix B $\in R^{f_g\times f_d}$ and a constant $\tau > 0$,, we have

$$D_\tau(B) = \underset{A\in R^{f_g\times f_d}}{\operatorname{argmin}} \tau\|A\|_* + \frac{1}{2}\|A - B\|_F^2. \tag{16}$$

Therefore, an iterative update of Z can be transformed into

$$Z^{k+1} = D_\delta\left(Z^k - \delta\nabla F_2\left(Z^k\right)\right). \tag{17}$$

Further, Theorem 3.4 of Reference [33] shows that if the minimum of the optimization problem (13) exists and $0 < \delta < 2/L_f$, then for any initial parameter Z_0, the solution sequence (14) converges to the minimum value of Equation (13), where L_f is the Lipschitz of the function $F_2(Z)$ continuous gradient, that is, for a convex function $F(X)$, $\exists L_f > 0$; for $\forall X_1, X_2$, the following inequality holds

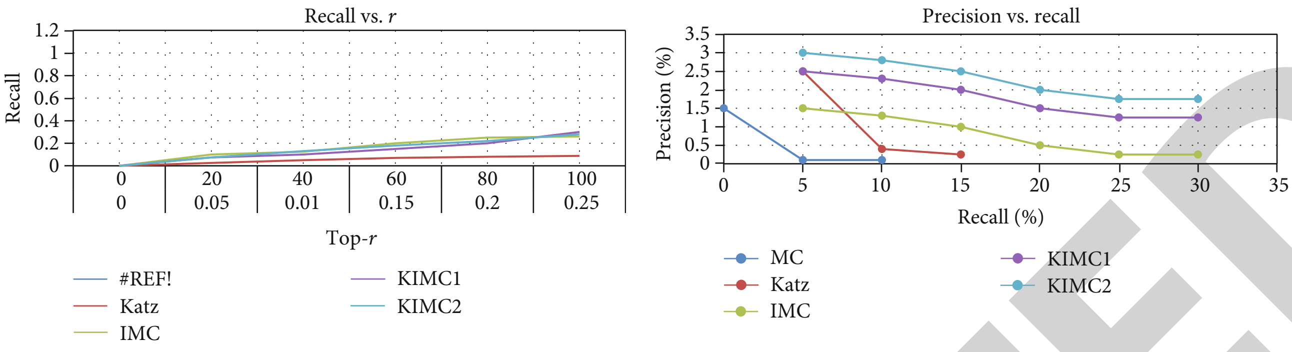
$$\|\nabla F(X_2) - \nabla F(X_1)\|_F \le L_f\|X_2 - X_1\|_F. \tag{18}$$

According to Proposition 1, if one can find a constant $L_f > 0$ and if $F_2(Z)$ satisfies Equation (18), then, the solution sequence (14) converges, and then, KIMC calculates the method combines, according to the Reference [33], Lemma 1, it is proved that

$$\begin{aligned} \|\nabla F(X_2) - \nabla F(X_1)\|_F^2 &= \left\|\lambda\Delta Z + \rho\left(X^T X\Delta ZY^T Y\right\|_F^2 \\ &\le 2\|\lambda\Delta Z\|_F^2 + 2\left\|\rho X^T X\Delta ZY^T Y\right\|_F^2 \\ &\le 2\lambda^2\|\Delta Z\|_F^2 + 2\rho^2\left\|\rho X^T X\Delta ZY^T Y\right\|_F^2 \\ &\le 2\lambda^2\|\Delta Z\|_F^2 + 2\rho^2\sigma_{\max}^2\left(X^T X\right)\sigma_{\max}^2\left(Y^T Y\right)\|\Delta Z\|_F^2 \\ &\le 2\lambda^2 + 2\rho^2\sigma_{\max}^2\left(X^T X\right)\sigma_{\max}^2\left(Y^T Y\right)\|\Delta Z\|_F^2. \end{aligned} \tag{19}$$

Therefore, the Lipschitz constant is

$$L_f = \sqrt{2\lambda^2 + 2\rho^2\sigma_{\max}^2\left(X^T X\right)\sigma_{\max}^2\left(Y^T Y\right)}. \tag{20}$$



(a)

(b)

Figure 3: (a) Overall performance w.r.t. various thresholds. (b) Overall performance w.r.t. various thresholds.

Table 2: Precision vs. recall %.

Precision %	Recall %	MC	Katz	IMC	KIMC1	KIMC2
0	0	1.5				
0.5	5	0.1	2.5	1.5	2.5	3
1	10	0.1	0.4	1.3	2.3	2.8
1.5	15		0.25	1	2	2.5
2	20			0.5	1.5	2
2.5	25			0.25	1.25	1.75
3	30			0.25	1.25	1.75

Table 3: Recall vs. r for new gene.

Recall	Top-r	Katz	IMC	KIMC1	KIMC2
0	0	0	0	0	0
0.05	20	0.025	0.1	0.075	0.135
0.01	40	0.05	0.125	0.1	0.16
0.15	60	0.65	0.725	0.7	0.76
0.2	80	0.85	0.925	0.9	0.96
0.25	100	0.85	0.92	0.9	0.96

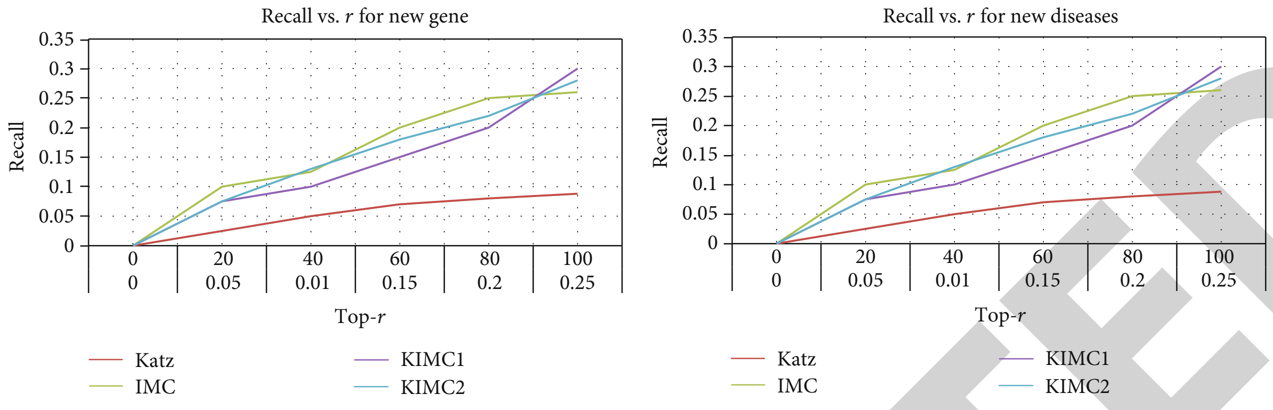
In this paper, the KIMC model with and without the regularization term of the elastic net is denoted as KIMC1 and KIMC2, respectively, and the solution process of KIMC2 is shown in Algorithm 1.

Katz status calculations are made possible for very large networks by an algorithm, although it should be noted that the metric has limited application. The Katz score is a modification of degree centrality, where distant players are taken into consideration through additional geometric series iterations. In fact, the Katz score frequently has a strong correlation with degree, offering a local gauge of centrality (based more on a node's immediate surroundings than its position across the larger network). So, even though other shortest-path or eigenvector centrality metrics offer a more comprehensive perspective, Katz scores nevertheless allow for the differentiation of actors of the same degree.

4. Experimental Results and Analysis

In this section, the gene-disease datasets and the sources of gene and disease characteristics used in the experiments are introduced, and the general evaluation criteria for gene-disease association prediction and the experimental results are analyzed in detail. Finally, the performance of several methods is compared.

4.1. Datasets and Features. The gene and disease information used in this study comes from the OMIM database, which not only includes relevant information on all monogenic diseases inherited in Mendelian fashion but also includes information on chromosomal diseases, polygenic diseases, and mitochondrial diseases, covering various conditions. Additionally, it gives details on information on the chromosomal location, linkage relationship, structure, and function of known pathogenic genes and describes the clinical knowledge of various genetic diseases. The data is updated in a timely and authoritative manner. The experiment uses the gene-disease dataset provided by the literature [15], which includes the gene-disease associations collected via OMIM dataset, including 12331 and 3209 genes and diseases, respectively, and a total of 3954 known genes-disease associations and gene-gene similarity information for 12331 genes and phenotype-phenotype similarity data (i.e., disease-disease similarity data) for 3209 diseases. In addition, the gene and disease signatures required in this study can be extracted from different types of biological data from other sources. For example, gene signatures are removed from gene microarray data, gene function interaction data, homologous gene-phenotype data of different species, disease similarity networks, clinical manifestation data of diseases, and analysis of a large number of medical literature data. Disease characteristics were obtained from the data. Faced with such complex data, principal component analysis (PCA) is usually utilized for dimensionality reduction to extract the main features of genes and diseases.



(a)

(b)

Figure 4: (a) Recall w.r.t. various threshold r. (b) Recall w.r.t. various threshold r.

Table 4: Recall vs. r for new diseases.

Recall	Top-r	Katz	IMC	KIMC1	KIMC2
0	0	0	0	0	0
0.05	20	0.025	0.1	0.075	0.075
0.01	40	0.05	0.125	0.1	0.13
0.15	60	0.07	0.2	0.15	0.18
0.2	80	0.08	0.25	0.2	0.22
0.25	100	0.088	0.26	0.3	0.28

However, this experiment uses the genes and disease features provided by the literature [11].

In multivariate data analysis, principal component analysis (PCA) is frequently employed to minimize the dimension of the data, facilitate further analysis, and enable efficient data summarization. It is now a helpful tool for analyzing microarray data. It is sometimes challenging to assess the overall gene expression differences between data from various groups or to categorize based on a very high number of genes for a specific microarray dataset. This study provides a gene selection technique based on Krzanowski's plan. Using data on cancer gene expression, we show how successful this method is and contrast it with several different gene selection methods. The optimal gene subset for maintaining the original data structure is chosen using the suggested strategy.

4.2. Evaluation Indicators and Methods. As with the Katz [15], MC [11], and IMC [11] methods mentioned above, the experiments are evaluated using 3-fold cross-validation. When considering the prediction performance, the top-r sorting method is used (that is, the gene score value corresponding to each disease column in the prediction result is sorted from large to small, and first r genes are taken as the candidate pathogenic genes of the respective disease and the other few gene-disease association prediction methods. When evaluating the performance in different ways, take the disease-related causative genes corresponding to different thresholds r, compare the known associations recorded in the test set, and compare each method's recall. They are calculated as follows:

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{21}$$

At the same time, the accuracy of the experimental results needs to be analyzed. It is calculated as follows:

$$\text{Precision} = \frac{TP}{TP + FN}. \tag{22}$$

Among them, TP represents the number of correctly identified associations in the known gene-disease associations in the test set, FN represents the number of associations that are not accurately determined in the known gene-disease associations in the test set, and FP represents the unknown gene-disease associations determined as associated quantities. In the current field of biological research, it is hoped to get a better prediction effect in a low threshold range, usually $r \leq 100$. Second, while evaluating the global performance of prediction methods, researchers pay more attention to new genes and diseases with research gaps than some widely studied genes and diseases, hoping to continuously discover valuable new genes and disease associations to promote the development of medical research. Therefore, various methods are also concerned here for novel genes that have only one known association but no association at training time and new diseases that have only one available association but not at training time. At the same time, to verify the effectiveness of the proposed method, the top 10 candidate genes of 8 common diseases were selected and compared with the database and literature reports.

4.3. Global Performance. Some recently proposed gene-disease association prediction methods were undertaken for comparison during experimentation, namely, MC, IMC, and Katz. The recall results of 3-fold cross-validation are shown in Figure 2 and Table 1, where the abscissa

Table 5: Prediction of top 10 candidates.

Leukemia MIM : 601626	Alzheimer's disease MIM : 104300	Insulin resistance MIM : 125853	Prostate cancer MIM : 176807
TP53 (7157) [11]	PSEN1 (5563) [4]	SHH (6469)	SHH (6469) [1]
PAX6 (5080)	PSEN2 (5564) [4]	FGFR2 (2263)	BMP2 (650) [18]
PITX2 (5308)	LFNG (3955)	STAR (6770)	IHH (3549)
PTEN (5728) [15]	MESP2 (145873)	DLK1 (8788)	DHH (50846)
RUNX2 (860)	DLL3 (10683)	FGF10 (2255)	SOX2 (6657)
FGFR3 (2261)	TCF15 (6939)	CYP11A1 (1583)	LMNA (4000)
FOXE3 (2301)	CIT (11113)	CYP11B1 (1584)	AKT1 (207) [16]
SPI1 (6688)	NKX3-2 (579)	CYP17A1 (1586)	SIX1 (6495)
TGFB2 (7042)	CHUK (1147)	LBX1 (10660)	IGF1R (3480)
CREBBP(1387)	ROR2 (4920)	HSD3B2 (3284)	STAT3 (6774) [7]
Schizophrenia	Breast cancer	Gastric cancer	Colorectal cancer
MIM : 181500	MIM : 114480	MIM : 137215	MIM : 114500
PSEN1 (5663)	TP53 (7157)	TP53 (7157)	TP53 (7157)
WNT4 (54361)	APC (324)	APC (324)	APC (324)
FGFR3 (2261)	CTNNB1 (1499)	AXIN1 (8312)	CTNNB1 (1499)
PITX2 (5038)	AXIN1 (8312)	KIT (3815)	AXIN1 (8312)
PAX3(5077)	FGFR3(2261)	KRAS (3845) [3]	BMPR1A (657)
MSX2(4488)	MSH2(4436)	MSH2 (4436)	FGFR3 (2261)
PAX2(5076)	CDKN2A(1029)	CTNNB1 (1499) [13]	BMPR1B (658)
PTEN(5728)	BRCA1(672) [2]	MSH6 (2956)	PTEN (5728)
TBX3(6929)	RAD51(5888) [2]	RAD51 (5888)	MSH2 (4436)
IHH(3549)	KRAS(3845)	BRCA1 (672)	SMAD4 (4089)

represents the value of different thresholds r, and the ordinate represents the recall. The performance of the proposed KIMC1 method and KIMC2 method is better than several other comparison methods when taking different threshold r. When the threshold is set to $r = 100$, the recall rates of several methods are 6.7% for the MC method, 11.3% for the Katz method, 23.2% for the IMC method, 26.5% for the KIMC1 method, and 27.6% for the KIMC2 method. The KIMC2 method with elastic net regularization has a specific improvement compared with the previously proposed IMC method of integrating genetic disease features. The proposed method combines the advantages of the Katz method and the inductive matrix completion method simultaneously, and the overall performance has been improved further. At the same time, it can be seen from the figure that adding elastic net regularization can effectively alleviate the influence of data noise and enhance prediction: effect and stability. Secondly, the precision-recall curves of the experimental results are also given here. As shown in Figure 3(b) and Table 2, the abscissa is the recall, and the ordinate is the precision. It can be observed from the figure that when the recall rate is greater than 4%, under the same precision rate, the recall rates of KIMC1 and KIMC2 are improved compared with the other three methods. The curves under different thresholds with and without elastic net regularization are also compared here. It can be found that the precision rate of KIMC2 after adding elastic net regularization is also significantly improved compared with KIMC1.

4.4. Prediction of New Genes and New Diseases. In gene-disease association prediction, there is often a problem that is easily overlooked; most of the genes and diseases recorded in the existing databases are genes and conditions with high recognition and association, and only a few are associated with a single gene; therefore, in the experimental evaluation, such genes and diseases with higher credit and association are often more likely to be predicted, while in reality, researchers pay more attention to those genes and conditions that are in the blank of research. Therefore, we only focus on those genes and needs that are known to be associated with a single association and hide these known associations during training to show the predictive power of different methods for new genes and diseases.

Within the range of the threshold $r \leq 100$, the recall rate of new genes is shown in Figure 3(a) and Table 3, the abscissa represents different thresholds, and the ordinate represents the recall rate of new genes. When the threshold range is $0 < r \leq 45$, the Katz method predicts better than IMC when using the gene-gene similarity network and disease-disease similarity network as auxiliary information. Because in a heterogeneous network, such data can more directly reflect the association between genes and diseases. However, using different gene and disease data extraction features, IMC performed poorly within this threshold range. When $r > 45$, the prediction effect of the IMC method is significantly improved, and the advantage of using feature information for prediction is reflected. The proposed

KIMC1 method and KIMC2 method integrate the benefits of the Katz and IMC methods, making the prediction performance more stable in different threshold ranges while improving the prediction efficiency. When $r = 100$, the new gene recall rate of the KIMC2 method was 17.4%. The recall rate of new diseases is shown in Figure 4(b) and Table 4, where the abscissa represents the threshold, and the ordinate represents the recall rate of new diseases. It can be found from the figure that the predictive ability of the KIMC1 method and KIMC2 for new illnesses is also better than several other comparison methods.

4.5. Prediction of Top 10 Causative Genes for Some Common Diseases. The above analysis of the predictive power of new genes is only verified on the known gene-disease association datasets in the OMIM database, and some disease-causing genes that are not recorded in the database cannot be evaluated and verified the overall effect will be low. Simultaneous association between genes also cannot be analyzed. Here, the top 10 pathogenic gene prediction results of several common diseases in real life are selected for analysis, and the effect of the proposed method is further explained. Eight common diseases are chosen here, namely, leukemia, Alzheimer's disease, insulin resistance, prostate cancer, schizophrenia, breast cancer, stomach cancer (gastric cancer), and colon cancer (colorectal cancer). During the experiment, all the relevant pathogenic gene information of these 8 diseases in the training data was hidden (that is, the columns corresponding to these 8 common diseases were all set to "0"), and the top 10 candidate pathogenic genes of the predicted diseases were shown in Table 5. In the table, the number after the infection (such as MIM:601626) represents its corresponding number in the OMIM database, and the number after the gene (such as PAX6 (5080)) represents the corresponding number of the gene in the NCBI database. The gene order in the table is arranged in descending order of the prediction score. Through the analysis of the candidate disease-causing genes in the table, it can be found that the disease-causing genes predicted by this method are not limited to the genes recorded in the gene-disease relationship dataset but also predict some disease-related genes discovered in later studies. For example, the genes associated with Alzheimer's disease include PSEN1 and PSEN2 [4]. These confirmed disease-related genes are shown in the table.

Secondly, it can be found from the table that there is a high degree of overlap between the top 10 predicted genes of these 8 diseases, and some genes are shared in the 8 diseases, such as TP53, KRAS, and RAD51, which have been confirmed to be associated with multiple cancers. They are closely related, so this is solid evidence to support the idea that these common genes represent etiological relationships between various diseases. That is, such shared genes can lead to a variety of conditions. Through the analysis of such shared genes, it is further verified that the prediction results of the KIMC method can show some commonalities of genes. Therefore, the KIMC method can provide a valuable reference for researchers to discover disease-causing genes and study the association between disease-causing genes. The Cancer Genome Atlas (TCGA) has changed our understanding of cancer, established the significance of cancer genomics, and even started to alter how the disease is handled in clinical settings. The effects extend even deeper, touching computational biology, health and scientific technology, and other study areas. Over a 12-year span, the Cancer Genome Atlas (TCGA) acquired, identified, and examined cancer samples from over 11,000 people. The procedure was intricate and continuously altering to take into account new technologies, the subtle differences between various cancer forms, and other shifting elements. In order to overcome the problem of data sparsity, the Katz technique was created, which built a gene-disease heterogeneous network by integrating information on intergene similarity, information on interdisease similarity, and information on gene-disease association information prediction. However, this approach cannot reliably forecast.

5. Conclusion

This paper proposes an enhanced inductive matrix completion based on the Katz gene-disease association prediction algorithm for the (KIMC) model. The algorithm combines the advantages of the Katz and IMC methods, which can effectively alleviate the impact of the PU problem encountered. First, in the face of highly sparse gene-disease association data, it can effectively help the data sparsity problem encountered by existing methods. Secondly, by introducing elastic net regularization, the influence of data noise is alleviated, and the noise tolerance of the algorithm is enhanced while improving the prediction effect. Compared with the existing prediction methods, the prediction effect of the KIMC method is significantly improved. It can also effectively predict new genes and diseases that researchers are more concerned about. This method is of great significance for reducing research costs and helping researchers deeply study different diseases' causative genes and gene correlations. Based on the KIMC method proposed in this paper, future studies might take into account combining more diverse biological data sources, and they might investigate how to effectively extract from this biological data the features of genes and diseases with stronger association to help improve the prediction effect.

Data Availability

The data shall be made available on request.

Conflicts of Interest

The authors declare that they have no conflict of interest.

Acknowledgments

This research work is self-funded.

References

[1] J. Xu, L. Chen, H. Wu, and S. Zha, "One-step multi-view inductive matrix completion for gene-disease associations prediction," in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 914–920, Melbourne, Australia, October 2021.

[2] H. Wang, Y. Wei, M. Cao, M. Xu, W. Wu, and E. P. Xing, "Deep inductive matrix completion for biomedical interaction prediction," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 520–527, San Diego, CA, USA, November 2019.

[3] X. Zeng, Y. Lin, Y. He, L. Lv, X. Min, and A. Rodríguez-Patón, "Deep collaborative filtering for prediction of disease genes," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 5, pp. 1639–1647, 2020.

[4] X. Wang, Y. Gong, J. Yi, and W. Zhang, "Predicting gene-disease associations from the heterogeneous network using graph embedding," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 504–511, San Diego, CA, USA, November 2019.

[5] Y. Hu, A. Sharma, G. Dhiman, and M. Shabaz, "The identification nanoparticle sensor using back propagation neural network optimized by genetic algorithm," *Journal of Sensors*, vol. 2021, Article ID 7548329, 12 pages, 2021.

[6] R. Al-Dalky, K. Taha, D. Al Homouz, and M. Qasaimeh, "Applying Monte Carlo simulation to biomedical literature to approximate genetic network," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 3, pp. 494–504, 2016.

[7] X. Zeng, N. Ding, and Q. Zou, "Latent factor model with heterogeneous similarity regularization for predicting gene-disease associations," in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 682–687, Shenzhen, China, December 2016.

[8] P. Mitra, T. Pijnenburg, and V. Sazonau, "Discovering gene-disease associations with biomedical word embeddings," in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 163–170, Miami, FL, USA, December 2020.

[9] J. W. Luo, Y. Liu, P. Liu, Z. Lai, and H. Wu, "Data integration using tensor decomposition for the prediction of miRNA-disease associations," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 5, pp. 2370–2378, 2022.

[10] J.-Y. Shi, "The need of accelerators in analyzing biological networks," in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1436–1436, Shenzhen, China, December 2016.

[11] R. Li, Y. Dong, Q. Kuang et al., "Inductive matrix completion for predicting adverse drug reactions (ADRs) integrating drug-target interactions," *Chemometrics and Intelligent Laboratory Systems*, vol. 144, pp. 71–79, 2015.

[12] T. Thakur, I. Batra, M. Luthra et al., "Gene expression-assisted cancer prediction techniques," *Journal of Healthcare Engineering*, vol. 2021, Article ID 4242646, 9 pages, 2021.

[13] S. Sreekala and K. A. A. Nazeer, "A literature search tool for identifying disease-associated genes using hidden Markov model," in *2014 First International Conference on Computational Systems and Communications (ICCSC)*, pp. 90–94, Trivandrum, India, December 2014.

[14] C. H. Lee, O. Koyejo, and J. Ghosh, "Identifying candidate disease genes using a trace norm constrained bipartite raking model," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3459–3462, Osaka, Japan, July 2013.

[15] C. Fan, X. Lei, and F. X. Wu, "Prediction of CircRNA-disease associations using KATZ model based on heterogeneous networks," *International Journal of Biological Sciences*, vol. 14, no. 14, pp. 1950–1959, 2018.

[16] 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.

[17] 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, article e1000641, 2010.

[18] P. Jia, S. Zheng, J. Long, W. Zheng, and Z. Zhao, "dmGWAS: dense module searching for genome-wide association studies in protein–protein interaction networks," *Bioinformatics*, vol. 27, no. 1, pp. 95–102, 2011.

[19] L. Huang, L. Zhang, and X. Chen, "Updated review of advances in microRNAs and complex diseases: taxonomy, trends and challenges of computational models," *Briefings in Bioinformatics*, vol. 23, no. 5, 2022.

[20] L. Huang, L. Zhang, and X. Chen, "Updated review of advances in microRNAs and complex diseases: experimental results, databases, webservers and data fusion," *Briefings in Bioinformatics*, vol. 23, no. 6, article bbac397, 2022.

[21] X. Chen, C. C. Yan, X. Zhang, and Z. H. You, "Long non-coding RNAs and complex diseases: from experimental results to computational models," *Briefings in Bioinformatics*, vol. 18, no. 4, pp. 558–576, 2017.

[22] C. C. Wang, C. D. Han, Q. Zhao, and X. Chen, "Circular RNAs and complex diseases: from experimental results to computational models," *Briefings in Bioinformatics*, vol. 22, no. 6, article bbab286, 2021.

[23] C. Tang, H. Zhou, X. Zheng, Y. Zhang, and X. Sha, "Dual Laplacian regularized matrix completion for microRNA-disease associations prediction," *RNA Biology*, vol. 16, no. 5, pp. 601–611, 2019.

[24] X. Zheng, C. Zhang, and C. Wan, "miRNA-disease association prediction via non-negative matrix factorization based matrix completion," *Signal Processing*, vol. 190, article 108312, 2022.

[25] P. Alexiou, M. Maragkakis, G. L. Papadopoulos, M. Reczko, and A. G. Hatzigeorgiou, "Lost in translation: an assessment and perspective for computational microRNA target identification," *Bioinformatics*, vol. 25, no. 23, pp. 3049–3055, 2009.

[26] T. M. Witkos, E. Koscianska, and W. J. Krzyzosiak, "Practical aspects of microRNA target prediction," *Current Molecular Medicine*, vol. 11, no. 2, pp. 93–109, 2011.

[27] Human Protein Reaction Database, HPRDAugust 2012, http://www.hprd.org.

[28] 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.

[29] C. Deng, C.-X. Lin, and H.-D. Li, "Improving the prediction of disease-associated genes by integrating annotated gene sets," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 386–391, Houston, TX, USA, December 2021.

[30] B. Wang, X. Yao, Y. Jiang, C. Sun, and M. Shabaz, “Design of a real-time monitoring system for smoke and dust in thermal power plants based on improved genetic algorithm,” *Journal of Healthcare Engineering*, vol. 2021, Article ID 7212567, 10 pages, 2021.

[31] S. Ma, K. Yang, X. Zhou, X. Xu, and W. Liu, “Similarity-based algorithms for disease terminology mapping,” in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1378–1384, Shenzhen, China, December 2016.

[32] A. S. Sidhu, T. S. Dillon, E. Chang, and B. S. Sidhu, “Ontology-based knowledge representation for protein data,” in *INDIN '05. 2005 3rd IEEE International Conference on Industrial Informatics, 2005*, pp. 535–539, Perth, WA, Australia, August 2005.

[33] Z. Li, W. Zhang, R. S. Huang, and R. Kuang, “Learning a low-rank tensor of pharmacogenomic multi-relations from biomedical networks,” in *2019 IEEE International Conference on Data Mining (ICDM)*, pp. 409–418, Beijing, China, November 2019.