A Joint Weighted Nonnegative Matrix Factorization Model via Fusing Attribute Information for Link Prediction

Minghu Tang(✉)

School of Computer Science, Qinghai Minzu University, Xining 810007, China
mhtang@tju.edu.cn

Abstract. Link prediction is a widely studied problem and receives considerable attention in data mining and machine learning fields. How to efficiently predict missing or hidden edges in the network is a problem that link prediction needs to solve. Traditional link prediction only focuses on the information of network topology and ignores some non-topological information, which makes the prediction performance of algorithm decline rapidly when encountering extremely sparse network. To compensate for this deficiency, this paper proposes a joint weighted nonnegative matrix factorization model for link prediction via incorporates attribute information. By designing a weighted matrix to process the attribute information of each node, both the structure and attribute information fused into the nonnegative matrix factorization framework can fully play a guiding role in the link prediction task, thus solving the problem of structure sparsity and improving the prediction performance of the algorithm. Extensive experiments on five attribute networks demonstrate that the proposed model has better prediction performance than the dozen benchmark methods and the state-of-the-art link prediction algorithms.

Keywords: Link prediction · Nonnegative matrix factorization · Attribute networks

1 Introduction

Link prediction is a widely studied problem and receives considerable attention in data mining and machine learning in the past decades. It aims to infer a link which is not observed in current network or will arise in the future network [1–8]. The network object of link prediction research is a complex topology structure abstracted from real-world physical systems. In general, people observe the interactive system in the real-world, extract the entities in the system as the vertices, and the interaction relationship between entities as the edges, and construct a topological graph corresponding to the physical system, namely complex network model. Then, the network model is taken as the research object to explore some laws underlying the physical interaction system and simulate their evolution mechanism. However, due to the complexity of the real physical system, the extractive network models are often structurally incomplete. That is, there is

Y. Chenggang et al. (Eds.): MobiMedia 2022, LNICST 451, pp. 190–205, 2022.
https://doi.org/10.1007/978-3-031-23902-1_15

always a missing situation of real information about the system in the complex network obtained through observation. The purpose of link prediction is to infer the missing or possible relationships in the future through this abstract complex network model, and to further study the evolution mechanism of real physical systems [9].

Because the research of link prediction problem is of great significance for the development of economy and society, its results are widely used in all walks of life in the real society [3, 4]. For example, analyze the evolution mechanism of the network [9], study the drug targeting relationship in the field of bioinformatics [10], realize the personalized recommendation of scenic spots or recommend new friends in social network [4, 11], and identify criminals in the field of public security [12–15].

At present, with the development of mobile Internet network, the amount of social information increases rapidly. When this real interaction system is abstracted into complex networks, the corresponding number of vertices becomes extremely large. However, the interaction relationship between the nodes did not grow significantly with the node scale. This phenomenon leads to exist many vertices in the extracted complex networks, but the edges between them appear extremely sparse. The phenomenon that the number of links known in the network is much less than the number of no links is called the structural sparsity problem. This problem has a very large impact on the performance of the link prediction [16, 17]. Therefore, how to solve the problem of declining prediction performance due to structural sparsity in large-scale networks becomes a challenge for link prediction. The motivation of this paper is to study the fusion problem of node attributes, so as to dig out the auxiliary information that can compensate for the sparsity of network structure, and build a multi-source information integration mode to realize the improvement of link prediction performance.

Recently, Social platforms based on mobile Internet networks are very frequently used, many network datasets appear with both the topology and node attribute information. For example, a webpage (i.e., vertex) can be associated with other webpages via hyperlinks, and it may have some inherent attributes of itself, like the text description in the webpage. Such type of networks is known as attributed networks. Some studies have shown that the degradation of the link prediction performance due to the sparse structure can be alleviated to some extent by using the node attribute information [17]. Recently, some link prediction methods are proposed based on attribute networks [18–22]. However, due to the diversity and heterogeneity of information and the variability of fusion methods, these algorithms either have poor overall prediction, or lack sufficient migration and robustness, or have too high computational complexity to adapt to large-scale networks. Therefore, the problem of how to reasonably integrate the structure and node attribute information has largely not been successfully solved.

Non-negative matrix factorization (NMF) is an important technique in the field of machine learning [23]. It can integrate heterogeneous information and promote each factor information to play a potential role [24]. In general, for a given matrix $X \in R_+^{n\times m}$, the NMF algorithm tries to find two non-negative factor matrices $B \in R_+^{n\times k}$ and $C \in R_+^{k\times m}$, make $X = X' \approx BC$. Where the k is called internal rank or hidden space, it satisfies $(m+n)k \ll m$. The solution of NMF usually transforms into an optimization problem of finding $min_{B\geq 0, C\geq 0} L(X, BC)$, and the symbol $L(.,.)$ represents a certain loss function,

such as Euclidean distance, KL divergence, or IS divergence. Given the Euclidean distance, the above optimization method can be converted to $min_{B,C}||X - BC||_F^2$, and the $B \geq 0, C \geq 0$. Symbols $|| \cdot ||_F$ indicate the Frobenius norm. The F-norm of a general matrix is usually defined as $||X||_F = \sqrt{\sum_{ij} |x_{ij}|^2} = \sqrt{tr(X^T X)}$.

Considering the advantages of NMF models when incorporating multi-source information. In this paper, we introduce a *Joint Weighted Nonnegative Matrix Factorization* method for link prediction on attributed networks, namely JWNMF. For a given attributed network, our method presents a mechanism by using joint-NMF to integrate the structural and attribute information. Specifically, we design two matrix factorization terms. One is modeling the topology structure and the other is for attributes. Meanwhile, we modify the NMF by introducing a weighting variable for each attribute, which can be automatically updated and determined in each iteration.

Experiments are performed on five real-world attribute network datasets. The results show the advantages of performance of JWNMF model comparison with the benchmark methods and advanced algorithms.

The rest article develops as follows. Section 2 shows the related works. Section 3 is the network description and the problem definition. Section 4 is about the establishment of the proposed model and its optimization. Section 5 is experimental design and results analysis. The last part contains our conclusions and prospects.

2 Related Work

As a research hotspot in the field of complex network science, link prediction has been widely concerned by researchers in recent years. However, there are not much studies to fuse non-topological information like node attributes with network topological information and then realize link prediction, especially the framework based on NMF. Han et al. [16] used the configuration files of online social-contact users and other non-topological information, such as workplace and school to compute the attribute similarity, for counting the number of attributes the users all possess and the geographic distance between the users. Then, they proposed a prediction model based on support vector machines. Wang et al. [17] extracted topological and non-topological information by an implicit feature representation model, then proposed a link prediction method for missing link. Li et al. [18] proposed a link prediction for dynamic attributed networks. Moreover, for attribute networks with isolated nodes, the literature [17, 19–22] makes full use of attribute information to achieve link prediction on semi-structured networks.

However, it is difficult to integrate multi-source heterogeneous information and make them work in experimental prediction tasks. In this respect, the method based on matrix factorization is widely used [23, 24]. Menon et al. [25] proposed a link prediction algorithm based on the matrix factorization. Pech et al. [26] proposed a matrix filling-based link prediction method using the matrix filling principle in the field of recommendation systems. For the network topology sparsity, Chen et al. [27] proposed a link prediction model of robust NMF by using manifold regularization and sparse learning. To make full use of the node attribute information, Chen et al. [28] proposed a link prediction model incorporating node attribute information based on NMF, but the time complexity

of their algorithm is high. Jiao et al. [29] proposed a Link predication model based on matrix factorization. This model fused multi class organizations information of network. They take advantage of the auxiliary information beyond the node attributes. Chen et al. [30] proposed a novel link prediction model based on deep NMF, which elegantly fuses topology and sparsity-constrained to perform link prediction tasks. Inspired by the matrix perturbation principle, Wang et al. [31] proposed a perturbation-based model for NMF link prediction. Moreover, there are also some NMF-based prediction models, they are used in dynamic time-varying networks [32, 33].

3 Preliminary

In this section, we introduce the formalized description of the problem of link prediction, and the network definition.

3.1 Network Representation

Given an undirected attribute network $G(V, E, A)$ with n nodes, where $V = \{v_1, v_2, \cdots v_n\}$ is the set of nodes and $E = \left\{\left(v_i, v_j\right), 1 \leq i \leq n, 1 \leq j \leq n, i \neq j\right\}$ is the set of edges. The A is the set of attributes of all nodes in network. For the network G with n vertex, there are m attributes value for each vertex. These attributes are available to be represented by a matrix $A_{n\times m}$. Each row of the matrix $A_{n\times m}$ represents an attribute vector of the corresponding node v_i. If the node v_{i} has the k-th attribute value, then $A_{\mathrm{ik}} = 1$, otherwise $A_{\mathrm{ik}} = 0$. The topology structure of the attribute network is represented by an adjacency matrix $S_{n\times n}$. The element of the i^{th} row and the j^{th} column in the matrix correspond to the link between nodes v_i and v_j in the network, where $S_{ij} = 1$ if there is a link from v_i to v_j and $S_{ij} = 0$ otherwise. Multiple edges between two nodes and back edges on single nodes are not allowed.

3.2 Link Prediction Problem

The purpose of link prediction is to infer the probability P_{ij} of the existence of an edge between any two nodes v_i and v_j by using the known information in the network. In general, based on the sociological principle that "the more similar people are more likely to be connected", the P_{ij} is treated as some similarity between nodes v_i and v_j. The higher P_{ij}, the more similar v_i and v_j are, and the more likely v_i is to form a link with v_j. For a given observation network G, the P_{ij} probability of forming edges between unconnected nodes is inferred through the model proposed. The predicted values are then arranged in descending order, and the pairs of nodes at the top are considered the most likely to form connections. In this paper, we compute the score P_{xy} based on JWNMF model.

4 Proposed Method

In this section, we will introduce our proposed method in detail, which aims to fuse the attribute information of the nodes into the link prediction process.

4.1 Link Prediction Model: JWNMF

Excavating the available information and constructing a reasonable information fusion mode are the main ideas to solve the problem of network topology sparsity, and realize the link prediction task. Therefore, the basic framework of NMF is used to fully integrate the node attributes and network structure information to compensate for the defects of incomplete topological information, to realize the link prediction task and improve the performance in this paper. First, based on the basic principle of NMF, the adjacency matrix S representing the network topology is decomposed into the product of two non-negative factor matrices, namely $S \approx VV^T$, and the matrix A representing the node attribute information decomposed into $A \approx ZU^T$. However, the aim of this paper is to address information integration. Therefore, in order to enable the network structure and node attribute information to fully integrate and play a leading role in the link prediction, we need to attach certain constraint rules to their decomposed factor matrix. Inspired by the methods described in ref [24], which often delivers promising results for graph clustering, we apply the idea for attributed graph link prediction. Here, the hidden space V after the network structure information S is decomposed is approximately equal to the hidden space Z of the node attribute information, so that it can remain the same in the process of model learning, so as to achieve the purpose of mutual fusion and constraining the network structure and node attribute information. Therefore, the partial information of the attribute A is decomposed into the hidden space V of the structure information, namely $A \approx VU^T$. When the two-source information is integrated in a unified framework and uses Euclidean distance as a loss function, the overall model framework for the link prediction task is expressed as follows:

$$L = min_{V,U} ||S - VV^T||_F^2 + \lambda ||A - VU^T||_F^2 s.t. V \geq 0, U \geq 0, \tag{1}$$

where $S \in R_+^{n \times n}$, $A \in R_+^{n \times m}$, the factor matrix $U \in R_+^{m \times k} and V \in R_+^{n \times k}$ represent the hidden space that integrates topological structure and node attribute information, R_+ represents non-negative real number sets. The parameter $\lambda > 0$ balance the availability of structure and attribute information.

Since the node attributes in the network are easy to mix with noise, in order to further reduce the impact of the noise on the prediction results, and promote the guiding role of the attribute information in predicting the network structure information, we also introduce a matrix W to assign a weight for each attribute. At this point, the decomposition form can be expressed as $AW \approx VU^T$. By assigning a weight information to each node attribute with the matrix W, the effect of similarity between the node attributes can be uniformly integrated into the structure information to provide a promotion for the final results of link prediction. The weight matrix W is set to a diagonal matrix, which satisfies $\sum_{i=1}^{m} W_{i,i} = 1$. After introducing the weight matrix W, the complete objective function is expressed as follows:

$$L = min_{V,U} ||S - VV^T||_F^2 + \lambda ||AW - VU^T||_F^2 s.t. V \geq 0, U \geq 0, \tag{2}$$

where, the weight matrix $W \epsilon R_{+}^{m \times m}$. To ensure that the W weights are assigned to a rule space, update operations need to be normalized to:

$$W = \frac{W}{\sum_{i=1}^{m} W_{i,i}}, \tag{3}$$

4.2 Update Rules

The solution of model L is difficult to obtain the global optimal solution, but the local optimal solution can be realized by the multiplicative iterative method. Therefore, for V, U and W three factor matrices, introduce their corresponding non-negative Lagrangian multiplier α, β, γ, thus replacing the objective function Eq. (2) with an unconstrained loss function form:

$$L = \frac{1}{2}\left(||S - VV^T||_F^2 + \lambda||AW - VU^T||_F^2\right) + Tr\left(\alpha^T V\right) + Tr\left(\beta^T U\right) + Tr\left(\gamma^T W\right), \tag{4}$$

Simplified the Eq. (4) and take the partial differentiations of L for V, U, W respectively, then

$$\frac{\partial L}{\partial V} = -\left(SV + S^T V + \lambda AWU\right) + 2VV^T V + \lambda VU^T U + \alpha, \tag{5}$$

$$\frac{\partial L}{\partial U} = -\lambda WA^T V + \lambda UV^T V + \beta, \tag{6}$$

$$\frac{\partial L}{\partial W} = -\lambda A^T VU^T + \lambda A^T AW + \gamma, \tag{7}$$

In this regard, according to complementary relaxation condition of the Karush-Kuhn-Tucker (KKT), we have $\alpha V = 0, \beta U = 0, \gamma W = 0$. Set $\frac{\partial L}{\partial V} = 0$, $\frac{\partial L}{\partial U} = 0$, $\frac{\partial L}{\partial W} = 0$, then the update rule for V, U, W is obtained.

$$V \leftarrow V\frac{SV + S^T V + \lambda AWU}{2VV^T V + \lambda VU^T U}, \tag{8}$$

$$U \leftarrow U\frac{WA^T V}{UV^T V}, \tag{9}$$

$$W \leftarrow W\frac{A^T VU^T}{A^T AW}, \tag{10}$$

The above update rules Eq. (8) - Eq. (10) can be solved by element value or by matrix form as a whole. During the model learning training, the three-factor matrix V, U, W is obtained based on the convergence condition of the objective function. Then, the approximate solution of original network topology structure is solved by using the decomposition formula $V \times V^T$. That is, after learning the matrix V through model training, we can obtain the similarity score between any two nodes in the network, or the

probability P_{ij} of exist edge between two nodes, and finally realize the link prediction task.

In general, in the learning process, the model will seek the local optimal solutions of V, U, W using the update rules. However, before implementing the update operation, the adjacency matrix S and the attribute matrix A need to be preprocessed and given an initial value.

$$S = \frac{S}{\sum_{i=1}^{n}\sum_{j=1}^{n} S_{i,j}}, \tag{11}$$

$$A = \frac{A}{\sum_{i=1}^{n}\sum_{j=1}^{m} A_{i,j}}, \tag{12}$$

Note that updates the weight matrix W use Eq. (3).

The model JWNMF integrates the network structure and node attribute information through the NMF framework, and assigns a weight constraint information to each node attribute through the introduced diagonal matrix W, so that the network structure and node attributes can maximize their respective roles in the model training and learning process to serve the final prediction results. A schematic diagram of the principle of the model JWNMF is shown in Fig. 1.

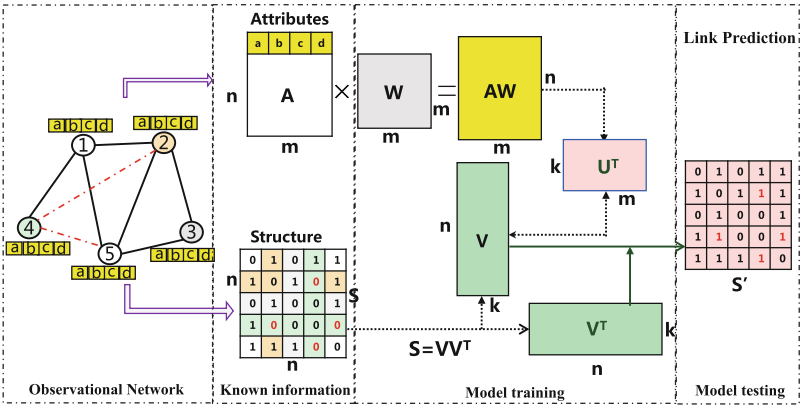

Fig. 1. A schematic diagram of the principle of JWNMF model.

In conclusion, according to the basic principles of the proposed JWNMF model, the pseudo-code description of the algorithm is designed as follows (shown in Table 1).

The experimental environment of this paper is based on the operating system of windows10 of x86 computer, and then the simulation experiment of link prediction is implemented with Matlab tool programming. Here, the computational complexity is discussed. The computational complexity of JWNMF algorithm comes mainly from the time cost when iteratively updating the matrix V, U, W. For a given network $G(V, E)$, the number of vertices V is n, and each vertex has m attributes. When updating V, U and W, to reduce the time overhead, we utilize the objective relative error as the stopping criterion and set to less than 10^{-6} in experiment. Moreover, the dimension

Table 1. Pseudo-code description of JWNMF algorithm

Algorithm Name: JWNMF
Input: S: the adjacency matrix of the given network, A: the auxiliary information matrix, k: number of features, λ: parameters.
Output: the approximate matrix of the network S
1: divide S into S^{train}, S^{test}
2: Initialize S and A by using Eq. (11) and Eq. (12).
3: Initialize V, U and W randomly.
4: do while
5: update V, U and W by means of Eq. (8) – Eq. (10).
6: get V after until object function convergence
7: end while
8: output VV^T

k after the matrix decomposition is a constant. Supposing the algorithm stops after t iterations, the overall cost for Symmetric NMF is $O(n^2kt)$. As the objective function adds one more linear matrix factorization term, the overall cost for updating rules is $O((n^2k + m^2k + mnk)t)$. According to the analysis rules of the time complexity of computer algorithms, when the scale n tends to infinity, the worst case of the time complexity of the model can be approximated by $O(n^2)$.

5 Experiment

This section mainly shows and analyzes the model prediction performance. Next, we will describe the datasets, comparison methods, evaluation metrics, and discussion of experimental results.

5.1 Datasets

This subsection mainly describes the basic topology of the datasets used in this paper, and the method of dividing training set and testing set.

To verify the model prediction performance, five real-world attribute network datasets widely used in the link prediction field were selected.

The basic topological properties of these network datasets are listed in Table 2. Where the symbol N represents the total number of network nodes, E represents the total number of existing links, $< K >$ is the network average degree, $< d >$ is the average shortest distance, C is the clustering coefficient, and *#attributes* represent dimension of node attributes. These network datasets used for the experiment can be downloaded from the following web sites. http://vladowiki.fmf.uni-lj.si/doku.php?id=pajek:data:urls:index; http://snap.stanford.edu/data/. For a detailed description of the data set, please also see the above website introduction.

Table 2. Topological information of the network datasets

Network	N	E	$< K >$	$< d >$	C	*#attributes*
Lazega	71	378	10.8	2.104	0.3853	7
Facebook	228	3419	29.991	1.868	0.6162	56
Cornell	195	286	2.903	3.2	0.1568	1703
Texas	187	298	3.027	3.036	0.1937	1703
Washington	230	366	3.373	2.995	0.1974	1703

5.2 Datasets Division Method

When comparing the prediction performance of the algorithm, the given network dataset needs to be divided according to the basic principles of machine learning. It is divided into training set and test set. There are many methods to divide data sets, and k-fold cross validation is used in this paper. The sample dataset was randomly divided into k parts, one of which was selected as test set and the remaining as training set, and then a prediction accuracy was calculated, so repeated k times. The prediction accuracy of the algorithm on the entire network dataset is the average of k prediction accuracies. In practical partitioning, k-taking 10 is a common method.

5.3 Evaluation Metrics

Like many existing link prediction studies, in our work adopts also the most frequently-used metrics AUC (area under the ROC curve) and the Precision to measure the performance of algorithm proposed. These metrics are viewed as a robust measure in the presence of data imbalance, which are also one of the most popular indices of evaluation link prediction. For more details on these two evaluation methods, readers can refer to the literature [1–4].

5.4 Baseline Methods

To validate the predictive performance of the newly designed algorithms, people usually select some benchmark methods and those representative up-to-date algorithms from the literature as the reference objects for comparative analysis. Generally, in order to reflect the fairness of comparison, the design principle of the comparison method selected is usually similar to the algorithm proposed. Therefore, in the experiment, several state-of-the-art algorithms based on NMF framework design often used in the link prediction research field are selected as reference objects. The benchmark methods are mainly structural similarity based classical algorithms.

We list four types of link prediction methods as the benchmark methods, including eleven local similarity indices based on the number of common neighbours between pairs of nodes (CN, AA, RA, PA, Salton, Jaccard, Sorenson, HPI, HDI, LHN and TSCN), four random walk methods (ACT, CosPlus, LRW, SRW), three local path methods (LocalP, Katz, LHN-II) and four other similarity algorithms (MFI, LNBCN, LNBAA, LNBRA).

The mathematical expressions of these methods and their definitions can be found in ref. [1–4].

In addition, four state-of-the-art NMF-based link prediction algorithms were used as comparison methods in the experiment. They are the original NMF method [23], matrix fill method (MC) [26], the attribute-based NMF method (NMF_LP) [28] and the NMF method based on the perturbation principle (SPM_NMF) [31] respectively.

5.5 Experimental Results and Analysis

This section provides a comprehensive analysis of the predictive performance of JWNMF model. Experiments were performed on five real-world network datasets. The prediction performance of various algorithms is fairly judged by two evaluation indicators, Precision and AUC, and show the final evaluation results. In experiment, the datasets were divided into test set E^{test} and training set E^{train} in different proportions. The results of experimental simulation are analyzed by taking the overall average at each proportion. Typically, the value of average prediction accuracy obtained from 100 independent simulations via Precision or AUC are taken as the final performance results.

5.6 Model Parameter Setting

To adjust the prediction performance of the JWNMF model to the optimum, the parameter λ in the model was analyzed before the start of the experiment. Figure 2 shows the predictive performance of the model when its parameter values are in the range from 0 to 3. Through the comparative analysis, the local optimum of the parameters λ is finally determined. In experiment, the parameter λ of JWNMF model takes a value of 0.09.

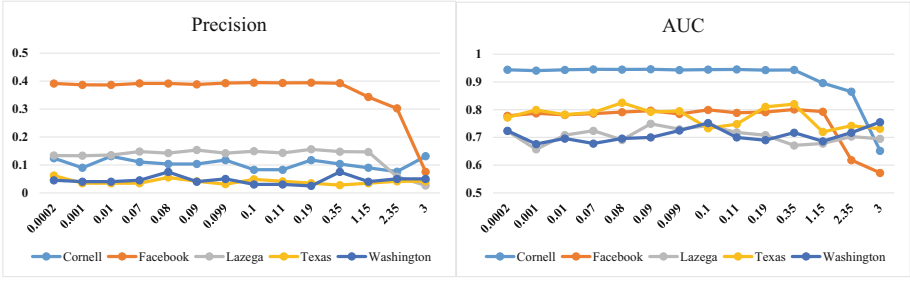

Fig.2. Analysis parameter λ of JWNMF model

5.7 Performance Analysis

According to the conventional way in the field of link prediction research, each experimental data set is first divided by the ratio of 20% to 90%, and the step size is 10, with a total of 8 proportions. Assuming that the total number of existent edges are $|E|=m$ in the

Table 3. Predictive results via AUC metrics on five network datasets

AUC	Texas	Cornell	Washington	Lazega	Facebook
CN	0.5449	0.5620	0.5637	0.6675	0.8863
AA	0.5596	0.5770	0.5695	0.6816	0.9017
RA	0.5562	0.5804	0.5698	0.6808	**0.9102**
PA	0.6970	0.7770	**0.7680**	0.6440	0.8350
Salton	0.5348	0.5435	0.5544	0.6562	0.8431
Jaccard	0.5283	0.5447	0.5518	0.6575	0.8579
Sorens	0.5280	0.5560	0.5150	0.6430	0.8280
HPI	0.5330	0.5670	0.5130	0.6490	0.7870
HDI	0.5270	0.5370	0.5220	0.6580	0.8230
LHN	0.5470	0.5370	0.5270	0.6350	0.7300
TSCN	0.5390	0.5530	0.5060	0.6800	0.4300
ACT	0.5977	0.5713	0.5983	0.6295	0.8476
CosPlus	0.5080	0.5540	0.4860	0.6660	0.9020
LRW_4	0.6500	0.6580	0.6720	0.7640	0.9100
SRW_3	0.6460	0.6250	0.6220	0.7200	0.9080
LocalP	0.5870	0.6110	0.6090	0.6600	0.9020
Katz	0.6539	0.6567	0.6935	0.7093	0.4389
LHNII.9	0.5017	0.5133	0.4910	0.5093	0.6380
MFI	0.6190	0.6720	0.6100	0.7040	0.8980
LNBCN	0.6070	0.6460	0.6290	0.6930	0.8730
LNBAA	0.5940	0.6680	0.6070	0.6730	0.9060
LNBRA	0.6230	0.6300	0.5900	0.6870	0.9090
NMF	0.5521	0.4950	0.4962	0.6783	0.8290
MC	0.5235	0.4470	0.4432	0.5000	0.5005
SPM_NMF	0.6260	0.7095	0.6362	0.7223	0.8745
NMF_LP	0.6421	0.7398	0.6705	0.7551	0.7795
JWNMF	**0.7080**	**0.8100**	0.7170	**0.7811**	0.8880

network. It indicates that 20% of the m are used for the training set when the partition ratio is 20%, while the remaining 80% is used as the test set.

The JWNMF model was trained on this training set together with the benchmark and contrast methods. To judge their prediction performance, the test set is then used to measure the effect. Each experiment needs to be run at least 100 times independently and then averaged as the result. Although the many results generated by experiments,

considering the universality and representativeness, Table 3 only shows the overall prediction effect of each method in the data set divided by 50%, the training set and the test set are in half each. The predictions values are shown in Table 3 by using AUC as the evaluation criterion.

From the numerical results presented in Table 3, The JWNMF method led the prediction performance on three datasets: Texas, Cornell and Lazega, but performed poorly on the Washington and Facebook datasets. As the overall analysis, the proposed JWNMF model showed good prediction performance on five datasets of attribute networks. This also shows that when implementing the link prediction, it can mine the external information such as the attributes of the nodes as an auxiliary, which can significantly improve the performance of the link prediction algorithm. This is significantly better effective than simply using structural information. Moreover, for networks with extremely sparse structure, using this external auxiliary information is more helpful to compensate for the insufficient performance problem caused by the sparse topological structure. Of course, the question of how to mine this auxiliary information and which external auxiliary information works better for the prediction is still under discussion. In order to show the overall predictive performance of the various methods more deeply, Fig. 3 shows the prediction effect when the data set is partitioned at 50% with Precision as the evaluation criterion.

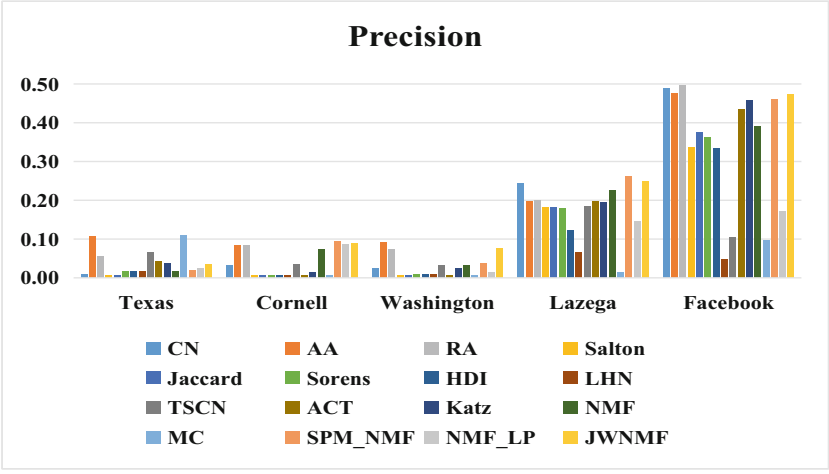

Fig. 3. Performance comparison based on the Precision metic

Above, we briefly mention that the network topology sparsity has obvious effects on the prediction performance of the algorithm. To demonstrate this problem more specifically, many experiments were deliberately designed and completed during the study. In these experiments, the network dataset was divided from dense to sparse in a ratio of 90% to 20%, and under each division scheme, the prediction performance of each algorithm is verified by Precision and AUC standards, to test the impact of network topology on the prediction results of each algorithm. Moreover, it is also used to verify

the adaptability and robustness of each algorithm under the different degrees of sparsity at network topology.

Taking the Facebook dataset as an example, the AUC values of each algorithm after the different partition proportions are shown in Table 4.

Table 4. The AUC value under different partitioning of Facebook dataset

AUC	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2
CN	0.9243	0.9237	0.9191	0.9069	0.8863	0.8642	0.8100	0.7041
AA	0.9355	0.9329	0.9267	0.9208	0.9017	0.8755	0.8227	0.7175
Salton	0.9260	0.9089	0.8954	0.8727	0.8431	0.8095	0.7620	0.6821
Jaccard	0.9067	0.9048	0.8927	0.8821	0.8579	0.8234	0.7674	0.6912
ACT	0.8468	0.8450	0.8462	0.8532	0.8476	0.8434	0.8344	0.8233
Katz	0.3394	0.3879	0.4147	0.3550	0.4389	0.4697	0.4002	0.2557
MC	0.8326	0.7954	0.7377	0.6458	0.5005	0.5000	0.5000	0.5000
NMF	0.9086	0.8879	0.8642	0.8527	0.8290	0.8004	0.7726	0.7419
SPM_NMF	0.9294	0.9158	0.9050	0.8907	0.8745	0.8575	0.8391	0.8059
JWNMF	0.9445	0.9369	0.9196	0.9073	0.8880	0.8614	0.8354	0.7853

From the analysis of these values, it can be seen that as the topology of the network gradually changes from dense to sparse, the prediction performance of the algorithm will have a significant downward trend. However, the prediction algorithm based on the JWNMF model still performs well at all proportions. This shows when facing different sparse degree of network topology, it can make full use of various external auxiliary information and compensate for the lack of topological information due to structure

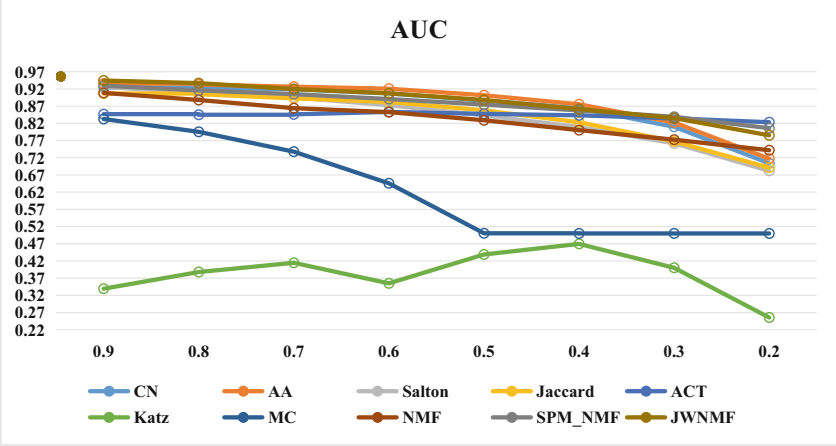

Fig. 4. The AUC value under different partition of Facebook dataset

sparsity. Thus, it basically ensures the prediction performance of the algorithm in abnormal cases, and improves the adaptability and robustness of the algorithm. Figure 4 shows this result more visually.

Similarly, with Precision as the evaluation criterion, we also compared the prediction performance of the various algorithms at different proportions of Facebook dataset (in Fig. 5). Although the model JWNMF is not the best under each partitioning scheme, it still shows a good prediction effect.

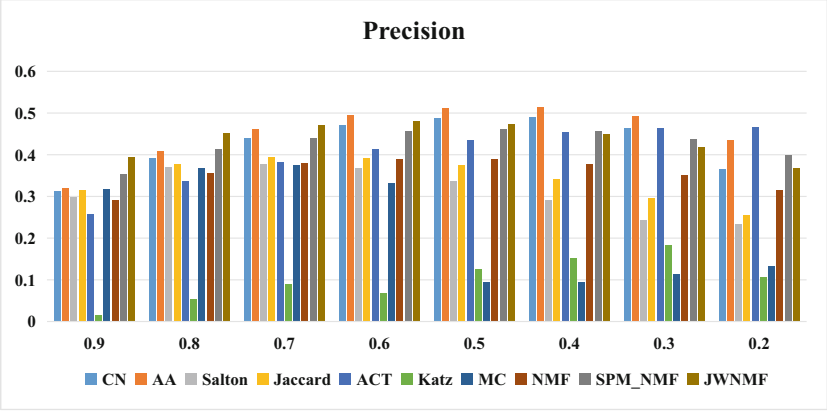

Fig. 5. The Precision value under different partition of Facebook dataset

6 Conclusion

In recent years, link prediction based on network topology has been one of the research hotspots in the field of data mining. However, in many cases, those algorithms that use only the information of network topology do not provide the accuracy required for link prediction, when the network topology is in an extremely sparse state. Furthermore, real-world networks are often sparse and contain noise, which makes the predictive performance of the algorithm very strongly correlated with the properties of the network itself. For these extremely sparse and noisy networks, the ultimate effect is not ideal if only the structural information is used to complete the prediction task. At present, with the development of mobile Internet, it is more and more convenient to obtain the non-topological information of network. This provides a hope for link prediction research.

In this paper, considering the advantages of NMF that is interpretability, nonnegative, and information fusion, we propose a link prediction model of weighted NMF. By designing a weighted matrix w to process the attribute information of each node, both the structure and attribute information fused into the NMF framework can fully play a guiding role in the link prediction task, thus solving the problem of structure sparsity and improving the prediction performance of the algorithm. Although our method can significantly improve the performance of link prediction on sparse networks, its temporal

complexity is relatively high. This is also a direction that we need to improve in the future. In addition, we also consider the cold-start link prediction of complex network in a semi-structured state as another target for future studies.

Acknowledgments. We would like to thank the anonymous reviewers for their contributions. This research was supported by the Teaching Reform Research Project of Qinghai Minzu University, China (2021-JYYB-009).

References

1. Martinez, V., Berzal, F., Cubero, J.C.: A survey of link prediction in complex networks. ACM Comput. Surv. **49**(4), 69–102 (2017)
2. Haghani, S., Keyvanpour, M.R.: A systemic analysis of link prediction in social network. Artif. Intell. Rev. **52**(3), 1961–1995 (2017). https://doi.org/10.1007/s10462-017-9590-2
3. Kumar, A., Singh, S.S., Singh, K., Biswas, B.: Link prediction techniques, applications, and performance: A survey. Phys. A **553**, 124289 (2020)
4. Daud, N.N., et al.: Applications of link prediction in social networks: A review. Journal of Network and Computer Applications **166**, 102716 (2020)
5. Rossi, A., et al.: Knowledge graph embedding for link prediction: A comparative analysis. ACM Trans. Knowl. Discov. Data **15**(2), 14–49 (2021)
6. Zhang, H.-F., et al.: Predicting missing links in complex networks via an extended local naïve Bayes model. EPL (Europhysics Letters) **130**(3), 38002 (2020)
7. Cai, L., et al.: Line graph neural networks for link prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
8. Singh, S.S., et al.: CLP-ID: Community-based link prediction using information diffusion. Inf. Sci. **514**, 402–433 (2020)
9. Zhang, Q.M., et al.: Measuring multiple evolution mechanisms of complex networks. Sci. Rep. **5**(1), 10350 (2015)
10. Nasiri, E., et al.: A novel link prediction algorithm for protein-protein interaction networks by attributed graph embedding. Computers in Biology and Medicine **137**, 104772 (2021)
11. Li, S., et al.: Friend recommendation for cross marketing in online brand community based on intelligent attention allocation link prediction algorithm. Expert Systems with Applications **139**(2020), 112839 (2020)
12. Bohannon, J.: Counterterrorism's new tool: "metanetwork" analysis. Science **325**(5939), 409–411 (2009)
13. Tayebi, M.A., Glässer, U.: Social network analysis in predictive policing: concepts, models and methods, pp. 7–14. Springer International Publishing (2016)
14. Assouli, N., Benahmed, K., Gasbaoui, B.: How to predict crime — informatics-inspired approach from link prediction. Physica A: Statistical Mechanics and its Applications, 570–125795 (2021)
15. Pang, G., et al. Deep learning for anomaly detection: A review. Association for Computing Machinery **54**(2) (2021)
16. Han, X., et al.: Link prediction for new users in social networks. IEEE International Conference on Communications (ICC), pp. 1250–1255 (2015)
17. Wang, Z., et al.: An Approach to Cold-start link prediction: establishing connections between non-topological and topological information. IEEE Trans. Knowl. Data Eng. **28**(11), 2857–2870 (2016)

18. Li, J., et al.: Streaming link prediction on dynamic attributed networks. Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, pp. 369–377 (2018)
19. Hao, Y., et al.: Inductive link prediction for nodes having only attribute information. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20), pp. 1209–1215 (2020)
20. Berahmand, K., Nasiri, E., Rostami, M., Forouzandeh, S.: A modified DeepWalk method for link prediction in attributed social network. Computing **103**(10), 2227–2249 (2021). https://doi.org/10.1007/s00607-021-00982-2
21. Shuo, Y., et al.: Inductive link prediction with interactive structure learning on attributed graph. ECML PKDD 2021: Machine Learning and Knowledge Discovery in Databases. Research Track, pp. 383–398 (2021)
22. Zhang, J.W., Kong, X.N., Yu, P.S.: Predicting social links for new users across aligned heterogeneous social networks. 2013 IEEE 13th International Conference on Data Mining (Icdm), pp. 1289–1294 (2013)
23. Gan, J., et al.: Non-negative matrix factorization: a survey. Comput. J. **64**(7), 1080–1092 (2021)
24. Kim, J., Shim, K., Cao, L., Lee, J.-G., Lin, X., Moon, Y.-S. (eds.): PAKDD 2017. LNCS (LNAI), vol. 10234. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-57454-7
25. Menon, A.K., Elkan, C.: Link prediction via matrix factorization. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 437–452. Springer (2011)
26. Pech, R., et al.: Link prediction via matrix completion. EPL (Europhysics Letters) **117**(3), 38002 (2017)
27. Chen, G., et al.: Robust non-negative matrix factorization for link prediction in complex networks using manifold regularization and sparse learning. Phys. A **539**, 122882 (2020)
28. Chen, B., et al.: Link prediction based on non-negative matrix factorization. PLoS ONE **12**(8), e0182968 (2017)
29. Jiao, P., Cai, F., Feng, Y., Wang, W.: Link predication based on matrix factorization by fusion of multi class organizations of the network. Scientific Reports **7**(1), 8937 (2017)
30. Chen, G., et al.: Link prediction by deep non-negative matrix factorization. Expert Systems with Applications **188**, 115991 (2022)
31. Wang, W., et al.: A perturbation-based framework for link prediction via non-negative matrix factorization. Scientific Reports **6**(10), 38938 (2016)
32. Zhang, T., et al.: Semi-supervised link prediction based on non-negative matrix factorization for temporal networks. Chaos, Solitons Fractals **145**, 110769 (2021)
33. Zhang, J., et al.: Temporal link prediction for cancer networks using structural consistency regularized non-negative matrix factorization. IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 280–283 (2021)