Research on Tacit Knowledge Discovery Methods Based on Two-mode Complex Networks: A Case Study of Potential Drug Target Mining (Post-print)

Authors: Li Dongqiao, Chen Fang, Han Tao, Yang Yanping, Wang Xuezhao, Wang Yanpeng, Cynthia Liu, Yingzhu Li

Date: 2023-04-01T16:16:03+00:00

Abstract

[Purpose/Significance] To reveal implicit knowledge hidden in massive literature by constructing a bipartite complex network model. [Methods/Process] The NetworkX complex network toolkit is utilized to construct a bipartite complex network model based on co-occurrence relationships between any two nodes; co-occurrence relationships of nodes in the network model are weighted, network topological information is calculated, and AP clustering is performed to extract direct relationships between nodes; the AUC method is employed to evaluate four link prediction algorithms—AA, JC, weighted improved wAA, and wJC—to select the most appropriate prediction algorithm and conduct predictive analysis of implicit relationships in complex networks. [Results/Conclusion] Empirical research using potential drug target mining as an example demonstrates that the wAA link prediction algorithm is optimal; the bipartite complex network model, metrics, and methodological system exhibit certain effectiveness for drug target mining within the Chemical Abstracts Service (CAS) database. Future plans include applying this model in other databases and research fields to further validate its generality and effectiveness.

Full Text

Research on Tacit Knowledge Discovery Methods Based on Two-Mode Complex Networks—Taking Potential Drug Target Mining as an Example

Authors: Li Dongqiao[1], Chen Fang[1], Han Tao[1], Yang Yanping[1], Wang Xuezhao[1], Wang Yanpeng[1], Cynthia Liu[2], Yingzhu Li[2]

[1]National Science Library, Chinese Academy of Sciences, Beijing 100190
[2]Chemical Abstracts Service, Columbus, OH 43202, USA

Abstract: [Purpose/Significance] This study aims to reveal tacit knowledge hidden in massive literature by constructing a two-mode complex network model. [Method/Process] Using the NetworkX complex network toolkit, a two-mode complex network model was constructed based on the co-occurrence relationships between any two nodes. Node co-occurrence relationships in the network model were weighted, network topology information was calculated, and AP clustering was performed to extract direct relationships between nodes. The AUC method was used to evaluate four link prediction algorithms—AA, JC, weighted improved wAA, and wJC—to select the most appropriate prediction algorithm and analyze hidden relationships in the complex network. [Result/Conclusion] An empirical study on potential drug target mining demonstrated that the wAA link prediction algorithm was optimal. The two-mode complex network model, indicators, and methodological system showed certain effectiveness in drug target mining within the Chemical Abstracts Service database. Future plans include testing in other databases or research fields to further verify the model's generality and effectiveness.

Keywords: tacit knowledge; link prediction; complex network; drug target; disease

1. Introduction

We live in an era of knowledge explosion characterized by big data, where the sheer volume of information has exceeded human capacity for knowledge absorption. As interdisciplinary research continues to intersect and integrate, studies in specific domains and topics have gradually attracted researchers' attention. Knowledge can be divided into explicit and tacit forms. Compared to explicit knowledge, tacit knowledge—due to its characteristics of being difficult to imitate and replicate—has become a key element for continuous researcher innovation. How to mine tacit knowledge hidden in massive literature represents both the foundation for researchers to maintain core competitive advantages and an inevitable opportunity and challenge in the current big data era.

Early tacit knowledge mining primarily relied on the "knowledge discovery model" known as the ABC model theory. Proposed by D.R. Swanson, an information science professor at the University of Chicago in 1987, the ABC model's fundamental principle is that for two unrelated literature sets A and C, if one set demonstrates that A can lead to B, while another shows that B can lead to C, then A and C must share a certain logical connection through this transitive relationship. Large-scale literature aggregation creates highly complex networks of relationships among research content, enabling researchers to mine relevant tacit knowledge through knowledge networks.

Link prediction, as a research direction in complex network data mining, primarily utilizes existing network information to predict relationships that exist but remain undiscovered, or relationships that currently do not exist but should or are likely to exist in the future. Current link prediction methods include neighbor-based approaches such as Adamic-Adar (AA) and Jaccard (JC), path-based methods such as Katz and FriendLink, and random walk-based methods such as Random Walk with Restart (RWR) and Local Random Walk (LRW).

Link prediction has achieved a series of successes in biomedical mining, primarily focusing on analyzing and mining associations between disease-disease, gene-disease, gene-protein, and protein-protein relationships from unstructured electronic medical record databases, UniProt experimental databases, Web of Science, PubMed, and other sources. However, these studies still have limitations in inferring associative relationships. First, data sources are relatively singular, with previous research mainly drawing from case databases, patent databases, or paper databases without integrated analysis of relevant database information, potentially leading to omitted predictions. Second, associated nodes are limited, as previous methodological approaches did not address relationships between homogeneous nodes in networks composed of two node types.

The Chemical Abstracts Service database is the world's largest integrator of chemical and related scientific information, containing not only paper and patent data but also indexed substance data such as genes, proteins, and drugs. Analyzing multi-source data can more comprehensively reveal tacit knowledge across an entire domain. Therefore, this study employs a two-mode complex network link prediction approach to analyze papers, patents, and substance data from the Chemical Abstracts Service, deeply mining tacit knowledge within and using potential drug target mining as an empirical case study. This approach provides valuable references for saving new drug development time and discovering additional potential drug indications.

2. Research Approach and Methods

This study proceeds through four aspects: "constructing a two-mode complex network model → extracting direct relationships in the complex network → selecting the optimal link prediction method → predicting hidden relationships in the complex network." The specific prediction route is shown in Figure 1 [Figure 1: see original paper].

2.1 Constructing the Two-Mode Complex Network Model

In complex networks, nodes represent different individuals in the real world, and edges represent relationships between individuals. When two different nodes share a specific relationship, they are connected by an edge; otherwise, no edge exists, where connected nodes are called adjacent nodes. Unlike regular and random networks, complex networks exhibit high complexity in three aspects: (1) structural complexity—connection structures are highly complex and may

change at any time; (2) node complexity—nodes may belong to multiple different types; and (3) various complexity factors mutually influence the network, with different factors producing different effects and potential connections between networks themselves.

A two-mode complex network is a representation mode of complex networks composed of two types of nodes. The two-mode complex network model can be expressed by the formula G = (T, D, L), where T and D represent any two associated nodes, and L represents the association relationship between any two nodes. The two-mode complex network used in this study differs from general bipartite networks. In bipartite networks, edges only connect different node types, with no edges between same-type nodes. In this study, L represents correlations extracted based on textual co-occurrence, meaning same-type nodes (e.g., D_1 and D_2, T_1 and T_2) necessarily have textual co-occurrence relationships that are important for hidden associations. Therefore, this study retains relationships between homogeneous nodes while introducing a parameter to assign different weights to edges connecting different node types.

2.2 Extracting Direct Relationships in the Two-Mode Complex Network

The basic properties of complex network topology are closely related to prediction method performance. Key properties include: network efficiency (the average of the sum of reciprocals of distances between all nodes), node degree (the number of nodes directly connected to a node in the network), average clustering coefficient (the average proportion of triangular structures containing any node), assortativity coefficient (the correlation between network degrees, measuring the tendency of nodes to connect), and average degree (the average of all node degrees in the network).

Direct relationship extraction in complex networks primarily uses SimRank similarity calculation on network graphs, followed by AP clustering based on results to extract known relationship features between nodes. SimRank similarity is based on the principle that if two nodes connect to similar nodes, then the two nodes are similar. It captures the overall graph structure information through recursive definition based on network topology. Compared to traditional text similarity, SimRank similarity calculation is entirely based on network graph topology, and its recursive definition enables the capture of global graph structure information. Unlike Google's PageRank algorithm, which only measures the importance of each node, SimRank similarity can compare the similarity between any two nodes, making it advantageous for calculating similarity matrices in this network.

The SimRank calculation formula is $S = C \cdot (W \cdot S \cdot W) + (1\text{-}C) \cdot I$, where S is the similarity matrix, W is the adjacency matrix, C is the decay factor, and I is the identity matrix. In this study's experiments, W is a weighted adjacency matrix with 20 iterations.

2.3 AP Clustering of Direct Relationships

AP clustering passes messages between nodes according to certain rules, generating cluster centers during multiple iterations to achieve automatic data point clustering. It offers advantages including fast clustering speed, no requirements for input similarity matrix triangle inequality or symmetry, and applicability to various scenarios. AP clustering's key advantage is eliminating the need to manually set initial cluster centers, instead relying on similarity matrix characteristics to gradually achieve clustering convergence. Based on SimRank-calculated similarity matrices, node vectors are constructed using the formula Node_{vec} = [S $_1$, S $_2$, … S , …, S], where i, j N.

The AP clustering algorithm from the sklearn toolkit was used with parameters: affinity='euclidean', convergence_{iter}=15, copy=True, damping=0.5, max_{iter}=200, preference=None, verbose=False. The AP clustering process automatically generates cluster numbers, initially exceeding 400 clusters. Before visualization, cluster centers were further iteratively clustered to obtain 28 clusters.

2.4 Selecting the Optimal Link Prediction Method

Link prediction is an important direction in complex network research, primarily using network topology information such as node degree, paths between node pairs, average shortest distance, and clustering coefficient to measure network similarity and predict the likelihood of connection formation between two nodes not yet connected. This study employs neighbor-based link prediction methods, introducing a weighted parameter α to perform weighted processing on AA and JC algorithms, where both homogeneous and heterogeneous nodes in the two-mode network are weighted.

The four algorithms are specifically formulated as follows:

(1) AA $= \Sigma_{\omega\, \Gamma(u)\, \Gamma(v)} \, 1/\log|\Gamma(\omega)|$

(2) wAA $= \Sigma_{\omega\, \Gamma(u)\, \Gamma(v)} \, (w(u,\omega)^{\wedge}\alpha + w(v,\omega)^{\wedge}\alpha) \,/\, \log(1 + w|\Gamma(\omega)|)$

(3) JC $= |\Gamma(u)\, \Gamma(v)| \,/\, |\Gamma(u)\, \Gamma(v)|$

(4) wJC $= w|\Gamma(u)\, \Gamma(v)| \,/\, |\Gamma(u)\, \Gamma(v)| = \Sigma_{\omega\, \Gamma(u)\, \Gamma(v)} \, (w(u,\omega)^{\wedge}\alpha + w(v,\omega)^{\wedge}\alpha) \,/\, (w|\Gamma(u)| + w|\Gamma(v)|)$

Where function $\Gamma(u)$ represents the neighbor nodes of u, $w|\Gamma(u)| = \Sigma_{\omega\, \Gamma(u)} w(w,x)^{\wedge}\alpha$, α is the weighting parameter ($\alpha = 1$ when (ω,x) (T,D); $\alpha = 0$ when (ω,x) (T,T) or (D,D)). The AA algorithm assigns different weights to different nodes in the common neighbor set, with each node's weight equal to the reciprocal of its degree logarithm. The wAA algorithm introduces weighting parameter α to the AA algorithm, where α reduces weights for same-type nodes while maintaining weights for different-type nodes proportional to original co-occurrence values. The JC algorithm represents node similarity as the proportion of common neighbor nodes to the total neighbor nodes of two nodes.

The wJC algorithm introduces weighting parameter α to the JC algorithm, with α functioning similarly to its role in wAA.

2.5 Evaluating Link Prediction Methods

This study uses the AUC (area under the receiver operating characteristic curve) method to evaluate the four link prediction algorithms: AA, JC, wAA, and wJC. AUC is the most commonly used standard for measuring link prediction algorithm accuracy, assessing overall algorithm precision. Ten-fold cross-validation was employed, randomly extracting 10% of connected node pairs and dividing them into ten parts (nine training sets and one test set) for prediction. Results from ten predictions were averaged to obtain AUC, calculated by the formula AUC = (n' + 0.5n'')/n, where n is the number of comparisons, n' is the number of times test set edge prediction values exceed non-existent edge values, and n'' is the number of times they are equal.

3. Experimental Results Analysis

With researchers' deepening understanding of disease mechanisms and continuous technological advancement, targeted drug therapy has played an increasingly important role, making target research a crucial direction in new drug development. This study constructs a target-disease two-mode complex network to predict other effective targets for disease treatment that have not yet been discovered, providing references for improving new drug development processes, saving research expenses, and reducing development risks.

3.1 Data Source and Processing

This study used the Chemical Abstracts Service database as the literature source, extracting 514,539 papers and patents related to antibody drugs and their contained antibody substances, with data acquisition ending in the first half of 2018. The relevant literature data was deeply indexed, and involved diseases, targets, substances, and other tags were manually cleaned and merged, forming a library of 1,015 antibody targets and 3,867 disease tags. Tumor diseases constituted the largest branch in the disease tag library, with 2,137 types (some tumor nodes overlapped with other disease categories, with duplicates removed from the total).

3.2 Constructing the Two-Mode Complex Network Model

This study used targets and diseases as nodes (named T and D , respectively), with their co-occurrence relationships in literature named L . Based on Python and the NetworkX toolkit, a two-mode complex network model was constructed and expressed by the formula G = (T , D , L). According to the data processing results in Section 3.1, T = 1,015, D = 3,867, and L = 911,479. Using the NetworkX toolkit, the target-disease network topology was analyzed. Based on

the co-occurrence frequency of any target T and disease D node pairs, relationships between nodes were weighted, and network topology metrics including node count, edges, efficiency, average clustering coefficient, weighted clustering coefficient, assortativity coefficient, and average degree were calculated, as shown in Table 1 .

Table 1 Basic topological properties of the target-disease two-mode complex network

Node count (T+D)	Average clustering coefficient	Weighted clustering coefficient	Efficiency	Assortativity coefficient	Average degree
4,882	0.3879	0.6661	6.1587$\times$10^{-4}	-0.1294	373.4

3.3 Extracting Direct Relationships in the Two-Mode Complex Network

SimRank similarity was used to calculate network graph topology information, and AP clustering was applied to cluster the similarity matrix, forming 28 clusters. Figure 2 [Figure 2: see original paper] shows the AP clustering diagram of direct relationships in the target-disease two-mode complex network, where each color represents a cluster. There were two giant clusters with more than 500 nodes: Cluster 10 (2,741 nodes) and Cluster 14 (563 nodes) (see Table 2).

Cluster 10 contained the most target-disease relationship pairs, including 672 target nodes and 2,069 disease nodes. The three targets with the most disease associations were epidermal growth factor receptor, carcinoembryonic antigen, and Notch ligand DLL4. The three diseases with the most target associations were Tendinitis, Cytopenia, and Primary sclerosing cholangitis. Cluster 14 ranked second in target-disease relationship pairs, containing 77 target nodes and 486 disease nodes. Its three most-connected targets were CD80 antigen, Ganglioside GD3, and Tumor-associated glycoprotein 72, while its three most-connected diseases were Autism, Congenital heart disease, and Severe combined immunodeficiency.

3.4 Selecting the Optimal Link Prediction Method

Table 3 summarizes the number of node pairs with prediction values > 0 and the counts of targets and diseases calculated by each prediction algorithm. The disease network represents the global disease network, while the tumor network represents a tumor disease subnetwork extracted after global network prediction. Weighted algorithms (wAA vs. AA, wJC vs. JC) produced identical counts, indicating weighted algorithms do not omit targets or diseases in statistics. However, differences existed between AA (wAA) and JC (wJC), with smaller differences in the disease network but more pronounced differences in the tumor

network. In the tumor network, AA (wAA) predicted more relationship pairs than JC (wJC) but involved fewer diseases.

Table 4 shows AUC evaluation results for the four algorithms. The wAA algorithm achieved the highest AUC value (0.9714), making it the optimal link prediction algorithm. Therefore, subsequent empirical analysis was based on wAA algorithm results.

Table 3 Node and relationship pair counts in prediction results

Algorithm	Predicted pairs (prediction value > 0)	Targets involved	Diseases involved
AA (wAA)	2,544,973	1,015	3,867
JC (wJC)	393,149	1,015	3,867

Table 4 Evaluation metrics for the two-mode complex network

Algorithm	AUC value
AA	0.9485
wAA	0.9714
JC	0.8826
wJC	0.9698

3.5 Predicting Hidden Relationships in the Two-Mode Complex Network

This study predicted over 2 million target-disease relationship pairs. Based on the principle that higher prediction values indicate greater likelihood of relationship existence, the NetworkX toolkit was used to filter the top 100 target-disease relationship pairs according to wAA algorithm results, with the target-disease complex network relationships visualized (see Figure 3 [Figure 3: see original paper]). In the figure, red circles represent targets, blue squares represent diseases, green solid lines represent direct relationships (known target-disease associations), and purple dashed lines represent hidden relationships (potential target-disease associations), with line thickness indicating relationship strength.

In the top 100 target-disease relationships, the five most closely connected direct relationship pairs were: (1) epidermal growth factor receptor and CD20 antigen, (2) epidermal growth factor receptor and vascular endothelial growth factor, (3) epidermal growth factor receptor and Tyrosine kinase receptor HER2, (4) CD20 antigen and Tyrosine kinase receptor HER2, and (5) CD20 antigen and Integrin αM.

Cytotoxic T-lymphocyte-associated protein 4 (CTLA4) and Programmed cell death protein 1 (PD-1) are two important targets for immune checkpoint therapy, enhancing specific anti-tumor immune responses. James P. Allison and Tasuku Honjo won the 2018 Nobel Prize in Physiology or Medicine for discovering immune checkpoint therapy targeting CTLA-4 and PD-1, respectively. Among the top 20 indirect target-disease relationships, CTLA4-targeted drugs for reactive arthritis had the highest prediction value, with additional predictions for angina pectoris, adult respiratory distress syndrome, and Raynaud disease, ranking 12th, 17th, and 19th respectively. PD-1-targeted drugs for reactive arthritis ranked 2nd, with predictions for allogeneic transplant rejection and connective tissue disease ranking 13th and 14th.

Vascular endothelial growth factor is also an important anti-tumor drug target. Table 5 shows that drugs targeting vascular endothelial growth factor had the most predicted disease indications, including Wiskott-Aldrich Syndrome, diphtheria, vaginitis, mouth disease, pertussis, and central nervous system inflammation.

Since the target-disease hidden relationship prediction did not show many target-tumor relationships, and tumors constitute a high proportion of the disease tag library, target-tumor hidden relationships were analyzed separately. Figure 4 [Figure 4: see original paper] shows the top 100 direct and hidden relationships between antibody targets and tumors. The three targets most directly related to tumors were CD20 antigen, Tyrosine kinase receptor HER2, and vascular endothelial growth factor.

Interleukin-1β is an important target for treating autoimmune diseases and Alzheimer's disease. Table 6 shows that Interleukin-1β-targeted drugs for liver neoplasm had the highest prediction value, with additional predictions for uterus neoplasm, small intestine neoplasm, ependymoma, and eye neoplasm, ranking 13th, 19th, and 20th respectively. Fc RI receptor is an important target for treating allergic diseases. Fc RI receptor-targeted drugs for astrocytoma ranked 2nd, with predictions for small intestine neoplasm, ependymoma, and eye neoplasm. Interleukin-5 is an important target for treating bronchial asthma, with Interleukin-5-targeted drugs predicted to treat the most disease types, including liver neoplasm, non-Hodgkin lymphoma, Hodgkin disease, glioma, metastasis, and testis neoplasm.

Table 5 Top 20 target-disease hidden relationship predictions

Target	Disease	Prediction value
Cytotoxic T-lymphocyte-associated protein 4	Reactive arthritis	0.252176

Target	Disease	Prediction value
Programmed cell death protein 1	Reactive arthritis	0.227286
Tyrosine kinase receptor HER2	Vaginitis	0.224872
Vascular endothelial growth factor	Wiskott-Aldrich Syndrome	0.224487
Interleukin-1β	Liver neoplasm	0.223732
Vascular endothelial growth factor	Diphtheria	0.223272
Fc RI receptor	Astrocytoma	0.220884
Interleukin-12 subunit β	Vaginitis	0.220090
Cytotoxic T-lymphocyte-associated protein 4	Liver neoplasm	0.219635
Programmed cell death protein 1	Hodgkin disease	0.217389
Programmed cell death protein 1	Vaginitis	0.216271
Vascular endothelial growth factor	Mouth disease	0.215814
Cytotoxic T-lymphocyte-associated protein 4	Angina pectoris	0.214442
Programmed cell death protein 1	Allotransplant rejection	0.211445
Cytotoxic T-lymphocyte-associated protein 4	Connective tissue disease	0.211154
Interleukin-6	Pertussis	0.209907
Cytotoxic T-lymphocyte-associated protein 4	Adult respiratory distress syndrome	0.208939

Target	Disease	Prediction value
Vascular endothelial growth factor	Central nervous system inflammation	0.208375
Cytotoxic T-lymphocyte-associated protein 4	Raynaud disease	0.206707

To further validate prediction results, we examined "CTLA4-targeted drugs for reactive arthritis, angina pectoris, adult respiratory distress syndrome, and Raynaud disease" as a case study. Using PubMed and Incopat patent databases as validation sources, we searched titles and abstracts with a timeframe through June 2018. Table 7 shows that before June 2018, these databases contained studies on CTLA4-related molecules or proteins in these diseases, but no papers or patents mentioned using CTLA4 as an antibody target for treating these specific conditions. These validation results demonstrate that the wAA link prediction algorithm can identify hidden knowledge associations not reported in existing publications, providing valuable references for researchers.

Table 7 Validation of target-disease hidden relationship predictions

Target	Predicted disease	PubMed search results	Incopat search results	Analysis
CTLA4	Reactive arthritis	2 comparative papers found on different binding proteins treating various diseases including reactive arthritis, but no studies using CTLA4 as antibody target	Similar findings	Prediction identifies unreported associations
CTLA4	Angina pectoris	3 papers on CTLA4-related molecules treating angina, but no antibody target studies	Similar findings	Prediction identifies unreported associations

Target	Predicted disease	PubMed search results	Incopat search results	Analysis
CTLA4	Adult respiratory distress syndrome	2 papers on syndrome mechanisms and CTLA4 drug use in a cancer patient with ARDS, but no targeted therapy studies	Similar findings	Prediction identifies unreported associations
CTLA4	Raynaud disease	1 paper on binding proteins treating Raynaud syndrome among other diseases, but no CTLA4 antibody target studies	Similar findings	Prediction identifies unreported associations

4. Conclusion

This study constructed a two-mode complex network model using the Python platform and NetworkX toolkit, employing the improved wAA link prediction algorithm for analysis. Using drug target prediction as an empirical case, the approach effectively revealed potential therapeutic targets in target-disease two-mode complex networks, providing references for saving new drug development time and discovering additional drug indications.

However, this research has limitations: (1) This study focused on neighbor-based link prediction algorithms without exploring path-based or random walk-based algorithms, which will be attempted in future work. (2) Due to hierarchical relationships among disease categories, whether to merge same-type diseases during prediction requires consideration. Merging may treat unknown relationships as known, omitting valuable detailed hidden associations, while not merging may treat some known relationships as unknown, introducing "noise" that requires expert manual review. How to effectively improve this requires further discussion. (3) This study empirically validated the two-mode complex network model in potential drug target mining; future work will test other research fields and databases to further verify the model's generality and effectiveness.

Acknowledgments

We thank Yi Deng, Ma Qingyang, Yu Min, and other personnel from Chemical Abstracts Service for their assistance and guidance in data analysis.

References

[1] Zhou Qingling. User tacit knowledge mining process and implementation technology [J]. China Science and Technology Information, 2015(11): 61-62.
[2] Lü Linyuan, Zhou Tao. Link Prediction [M]. Beijing: Higher Education Press, 2013.
[3] Lü Linyuan. Link prediction in complex networks [J]. Journal of University of Electronic Science and Technology of China, 2010, 39(5): 651-661.
[4] Yao Yabing. Research on link prediction methods based on complex network topology [D]. Lanzhou: Lanzhou University, 2017.
[5] Adamic LA, Adar E. Friends and neighbors on the Web [J]. Social Networks, 2003, 25(3): 211-230.
[6] Jaccard P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura [J]. Bulletin de la Société Vaudoise des Sciences Naturelles, 1901, 37: 547-579.
[7] Katz L. A new status index derived from sociometric analysis [J]. Psychometrika, 1953, 18(1): 39-43.
[8] Papadimitriou A, Symeonidis P, Manolopoulou Y. Fast and accurate link prediction in social networking systems [J]. Journal of Systems and Software, 2012, 85(9): 2119-2132.
[9] Brin S, Page L. Reprint of: The anatomy of a large-scale hypertextual Web search engine [J]. Computer Networks, 2012, 56(18): 3825-3833.
[10] Liu W, Lü L. Link prediction based on local random walk [J]. EPL, 2010, 89(5): 58007.
[11] Yu Huangyingzi, Dong Qingxing, Zhang Bin. Research on disease knowledge association mining and prediction methods based on network representation learning [J]. Information Studies: Theory & Application, 2019, 42(12): 156-162.
[12] Li Xing. Research on symptom-gene prediction methods based on complex networks [D]. Beijing: Beijing Jiaotong University, 2014.
[13] Bukek K, Mustafa P. Age-series based link prediction in evolving disease networks [J]. Computers in Biology and Medicine, 2015, 63: 1-10.
[14] Ding Liang. Research on non-coding RNA-disease correlation prediction based on heterogeneous network link prediction algorithms [D]. Anhui: University of Science and Technology of China, 2018.
[15] Hu H, Zhu CY, Ai HX. LPI-ETSLP: lncRNA-protein interaction prediction using eigenvalue transformation-based semi-supervised link prediction [J]. Molecular BioSystems, 2017, 13(9): 1781-1787.
[16] Wu Jinhua. Research on Alzheimer's disease protein networks based on data mining [D]. Shenyang: Liaoning University, 2018.
[17] Crichton G, Guo YF, Pyysalo S. Neural networks for link prediction in realistic biomedical graphs: a multi-dimensional evaluation of graph embedding-based approaches [J]. BMC Bioinformatics, 2018, 19: 176.
[18] Zhou Tao, Bai Wenjie, Wang Binghong, et al. Overview of complex network research [J]. Physics, 2005, 34(1): 31-36.
[19] Li Xing. Research on symptom-gene prediction methods based on complex

networks [D]. Beijing: Beijing Jiaotong University, 2014.
[20] Li Lanxi. Research on link prediction technology based on complex network structure [D]. Beijing: Beijing University of Posts and Telecommunications, 2019.
[21] Zhang Bin, Li Yating. Ranking robustness of link prediction results in collaboration networks [J]. Journal of Information Resources Management, 2018, 8(4): 89-97.
[22] Ge Jun. An overlapping community detection algorithm and its implementation on MapReduce [D]. Xi'an: Xidian University, 2013.
[23] Frey BJ, Dueck D. Clustering by passing messages between data points [J]. Science, 2007, 315(5814): 972-976.
[24] Wang Lin, Dong Xiaojiang. Parallel AP clustering method for community mining [J]. Microcomputer & Its Applications, 2017, 36(12): 16-18.
[25] Lü L, Zhou T. Link prediction in complex networks: A survey [J]. Physica A: Statistical Mechanics and its Applications, 2011, 390(6): 1150-1170.
[26] Yang Xiaocui, Song Jiaxiu, Zhang Xihuang. Link prediction algorithm based on network representation learning [J]. Computing Science and Exploration, 2019, 13(5): 812-821.
[27] Yang Yujie. Research on link prediction based on topological similarity in complex networks [D]. Beijing: Beijing University of Posts and Telecommunications, 2019.
[28] Chen Jiaying, Yu Jiong, Yang Xingyao, et al. Link prediction algorithm based on node importance in complex networks [J]. Journal of Computer Applications, 2016, 36(12): 3251-3255, 3268.

Author Contributions

Li Dongqiao: Conceptualized the research, processed data, performed statistical analysis, and wrote the manuscript.
Chen Fang: Participated in research design, implemented code, and organized research methods and data.
Han Tao: Supervised the research methodology.
Yang Yanping: Supervised the research methodology.
Wang Xuezhao: Supervised the research methodology.
Wang Yanpeng: Participated in methodology design.
Cynthia Liu: Participated in data processing.
Yingzhu Li: Participated in data processing.

Research on the Tacit Knowledge Discovery Based on Two-mode Complex Network—Taking Mining Potential Drug Targets as an Example

Li Dongqiao[1], Chen Fang[1], Han Tao[1], Yang Yanping[1], Wang Xuezhao[1], Wang Yanpeng[1], Cynthia Liu[2], Yingzhu Li[2]

[1]National Science Library, Chinese Academy of Sciences, Beijing 100190

[2]Chemical Abstracts Service, Columbus, OH 43202, USA

Abstract: [Purpose/Significance] This paper aims to extract tacit knowledge from massive literature by constructing a two-mode complex network model. [Method/Process] Through the NetworkX complex network toolkit, a two-mode complex network model was constructed based on the co-occurrence relationship of any two nodes. The co-occurrence relationship of nodes in the network model was weighted, the topology information of the network was calculated, and AP clustering was performed to extract the direct relationship between nodes. The AUC method was used to evaluate four link prediction algorithms—AA, JC, weighted improved wAA, and wJC—to select the most appropriate prediction algorithm and predict and analyze the hidden relationships of the complex network. [Result/Conclusion] The empirical study on potential drug target mining showed that the wAA link prediction algorithm was optimal. The two-mode complex network model, indicators, and method system were effective in drug target mining in the Chemical Abstracts Service database. The next step is to try other databases or research fields to further verify the generality and effectiveness of the model.

Keywords: tacit knowledge; link prediction; complex network; drug target; diseases

Note: Figure translations are in progress. See original paper for figures.