Collective Prediction of Multiple Types of Links in Heterogeneous Information Networks

Bokai Cao\textsuperscript{1}, Xiangnan Kong\textsuperscript{2}, Philip S. Yu\textsuperscript{1}

\textsuperscript{1}University of Illinois at Chicago
\textsuperscript{2}Worcester Polytechnic Institute
Heterogeneous Information Networks (HINs)

- Twitter social network
  - Nodes: users, tweets, locations
  - Links: follow, post, forward, check-in

- DBLP academic network
  - Nodes: authors, papers, journals
  - Links: authored by, cite, publish in
A Bioinformatics Heterogeneous Information Network Schema
## Applications of Link Prediction in Bioinformatics

<table>
<thead>
<tr>
<th>Predicted Link Type</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene —&gt; Gene</td>
<td>Protein-Protein Interaction Prediction</td>
</tr>
<tr>
<td>Gene —&gt; Disease</td>
<td>Prioritization of Candidate Disease Genes</td>
</tr>
<tr>
<td>Gene —&gt; Gene Ontology</td>
<td>Automated Gene Ontology Annotation</td>
</tr>
<tr>
<td>Chemical Compound —&gt; Gene</td>
<td>Drug-Target Interaction Prediction</td>
</tr>
<tr>
<td>Chemical Compound —&gt; Disease</td>
<td>Drug Discovery</td>
</tr>
<tr>
<td>Chemical Compound —&gt; Side Effect</td>
<td>Drug Side Effect Profiling</td>
</tr>
<tr>
<td>Chemical Compound —&gt; Chemical Ontology</td>
<td>Automatic Annotation of Chemical Ontology</td>
</tr>
</tbody>
</table>
A Bioinformatics Heterogeneous Information Network

Gene1 ! PPI ! Gene2 ! bind ! Drug1 ! treatDisease ! Drug3

PPI ! Gene3 ! causeDisease

Disease1 ! PPI ! Disease2 ! Disease3
Linkage Homophily Principle

- Conventional proximity measures for homogeneous links can not work on heterogeneous link prediction.

- Linkage Homophily Principle

  Two nodes are more likely to be directly linked if most of their respective similar nodes are linked.
Meta-path based Source Node Similarities

\[ Sim(s_i, s_j) \]

Generalized neighbors of source nodes

Gene \[\xrightarrow{\text{PPI}}\] Gene

Gene \[\xrightarrow{\text{bind}^{-1}}\] Drug \[\xrightarrow{\text{bind}}\] Gene

Gene \[\xrightarrow{\text{causeDisease}}\] Disease \[\xrightarrow{\text{causeDisease}^{-1}}\] Gene
Meta-path based Target Node Similarities

\[ \text{Sim}(t_p, t_q) \]

Generalized neighbors of target nodes

Disease \( \xrightarrow{\text{causeDisease}^{-1}} \) Gene \( \xrightarrow{\text{causeDisease}} \) Disease

Disease \( \xrightarrow{\text{treatDisease}^{-1}} \) Drug \( \xrightarrow{\text{treatDisease}} \) Disease
Node Similarity Extraction in HINs

Target node correlations

Source node correlations

meta-path

Heterogeneous Information Network
Relatedness Measure

Formulated based on the Linkage Homophily Principle:

\[ P(s_i, t_p) = \frac{\sum_{s_j \in S} \sum_{t_q \in T} Sim(s_i, s_j) \times I(s_j, t_q) \times Sim(t_p, t_q)}{\sum_{s_i \in S} \sum_{t_q \in T} Sim(s_i, s_j) \times Sim(t_p, t_q)} \]

Relatedness measure = \frac{\# existing links between generalized neighbors of \( s_i \) and \( t_p \)}{\# maximum potential links between generalized neighbors of \( s_i \) and \( t_p \)}

Weighted by the product of similarity between generalized neighbors and the node pair.
Multiple Types of Links in HINs

can imply
Link Prediction in HINs

- Conventional link prediction approaches assume that links are independent identically distributed (i.i.d.).

- In fact, links are interconnected via adjacent nodes in a network, thus are correlated with each other.
Heterogeneous Collective Link Prediction (HCLP)

- **Given:**
  - $L_i$: a set of existing links with type $i$, $i=1,\ldots,m$
  - $U_i$: a set of unlabeled links with type $i$, $i=1,\ldots,m$

- **Loop for $k$ iterations:**
  - **Loop $i$ for $m$ link types:**
    - Compute the relatedness measure from the HIN for $L_i$
    - Use $L_i$ to train a classifier $h_i$
    - Allow $h_i$ to label examples from $U_i$
    - Add these self-labeled examples with the most confidence into the HIN
Analogy to Co-training

- Different views share the same set of labels

HCLP: different types of links coexist in a HIN

- Exchange the most confident predictions on unlabeled data

HCLP: add the most confident predictions on each type of unlabeled links into the network
Experiments
SLAP Network Schema

#node types = 10
#link types = 11
#nodes > 290k
#links > 720k
# Compared Methods

<table>
<thead>
<tr>
<th></th>
<th>Independent link prediction</th>
<th>Collective link prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalized path count</td>
<td>ILP(NPC)</td>
<td>HCLP(NPC)</td>
</tr>
<tr>
<td>Random walk</td>
<td>ILP(RW)</td>
<td>HCLP(RW)</td>
</tr>
<tr>
<td>Symmetric random walk</td>
<td>ILP(SRW)</td>
<td>HCLP(SRW)</td>
</tr>
<tr>
<td>Relatedness measure</td>
<td>ILP(RM)</td>
<td>HCLP(RM)</td>
</tr>
</tbody>
</table>
Experimental Results

<table>
<thead>
<tr>
<th>Type of Links</th>
<th>ILP(NPC)</th>
<th>ILP(RW)</th>
<th>ILP(SRW)</th>
<th>ILP(RM)</th>
<th>HCLP(NPC)</th>
<th>HCLP(RW)</th>
<th>HCLP(SRW)</th>
<th>HCLP(RM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene $\xrightarrow{PPI}$ Gene</td>
<td>0.889 (5)</td>
<td>0.878 (8)</td>
<td>0.893 (3)</td>
<td>0.888 (6)</td>
<td>0.892 (4)</td>
<td>0.884 (7)</td>
<td>0.898 (2)</td>
<td>0.899 (1)</td>
</tr>
<tr>
<td>Gene $\xrightarrow{cause}$ Disease</td>
<td>0.690 (8)</td>
<td>0.696 (6)</td>
<td>0.701 (3)</td>
<td>0.697 (5)</td>
<td>0.696 (6)</td>
<td>0.701 (3)</td>
<td>0.708 (1)</td>
<td>0.704 (2)</td>
</tr>
<tr>
<td>Gene $\xrightarrow{has}$ Pathway</td>
<td>0.888 (7)</td>
<td>0.899 (5)</td>
<td>0.909 (2)</td>
<td>0.909 (2)</td>
<td>0.881 (8)</td>
<td>0.894 (6)</td>
<td>0.906 (4)</td>
<td>0.913 (1)</td>
</tr>
<tr>
<td>Gene $\xrightarrow{has}$ GO term</td>
<td>0.862 (5)</td>
<td>0.851 (8)</td>
<td>0.867 (4)</td>
<td>0.862 (6)</td>
<td>0.870 (3)</td>
<td>0.857 (7)</td>
<td>0.875 (1)</td>
<td>0.874 (2)</td>
</tr>
<tr>
<td>Gene $\xrightarrow{has}$ Gene family</td>
<td>0.669 (7)</td>
<td>0.674 (3)</td>
<td>0.673 (5)</td>
<td>0.678 (1)</td>
<td>0.669 (7)</td>
<td>0.673 (4)</td>
<td>0.677 (2)</td>
<td>0.672 (6)</td>
</tr>
<tr>
<td>Gene $\xrightarrow{has}$ Tissue</td>
<td>0.813 (8)</td>
<td>0.828 (6)</td>
<td>0.846 (3)</td>
<td>0.841 (4)</td>
<td>0.818 (7)</td>
<td>0.830 (5)</td>
<td>0.849 (2)</td>
<td>0.851 (1)</td>
</tr>
<tr>
<td>Chemical $\xrightarrow{bind}$ Gene</td>
<td>0.984 (4)</td>
<td>0.978 (7)</td>
<td>0.984 (4)</td>
<td>0.984 (4)</td>
<td>0.985 (2)</td>
<td>0.978 (7)</td>
<td>0.985 (2)</td>
<td>0.986 (1)</td>
</tr>
<tr>
<td>Chemical $\xrightarrow{treat}$ Disease</td>
<td>0.887 (4)</td>
<td>0.867 (8)</td>
<td>0.886 (5)</td>
<td>0.873 (7)</td>
<td>0.893 (3)</td>
<td>0.876 (6)</td>
<td>0.894 (2)</td>
<td>0.901 (1)</td>
</tr>
<tr>
<td>Chemical $\xrightarrow{cause}$ Side effect</td>
<td>0.901 (6)</td>
<td>0.886 (8)</td>
<td>0.913 (4)</td>
<td>0.931 (2)</td>
<td>0.913 (4)</td>
<td>0.890 (7)</td>
<td>0.927 (3)</td>
<td>0.939 (1)</td>
</tr>
<tr>
<td>Chemical $\xrightarrow{has}$ Substructure</td>
<td>0.934 (7)</td>
<td>0.938 (5)</td>
<td>0.943 (3)</td>
<td>0.939 (4)</td>
<td>0.935 (6)</td>
<td>0.932 (8)</td>
<td>0.951 (1)</td>
<td>0.947 (2)</td>
</tr>
<tr>
<td>Chemical $\xrightarrow{has}$ CO term</td>
<td>0.917 (6)</td>
<td>0.907 (8)</td>
<td>0.919 (5)</td>
<td>0.922 (3)</td>
<td>0.921 (4)</td>
<td>0.909 (7)</td>
<td>0.925 (2)</td>
<td>0.928 (1)</td>
</tr>
<tr>
<td><strong>Average Rank</strong></td>
<td>(6.1)</td>
<td>(6.5)</td>
<td>(3.7)</td>
<td>(4)</td>
<td>(4.9)</td>
<td>(6.1)</td>
<td>(2)</td>
<td>(1.7)</td>
</tr>
</tbody>
</table>
Robustness Analysis

- HCLP algorithm can effectively leverage complementary information from other link types, and improve the performances of predicting a particular link type even with its limited training data through a process of heterogeneous collective link prediction.

(a) Gene $\xrightarrow{PPI}$ Gene
(b) Chemical $\xrightarrow{bind}$ Gene
(c) Chemical $\xrightarrow{treat}$ Disease
Conclusions

- Studied the problem of **collective link prediction in HINs**:  
  - designed a **relatedness measure** between different types of objects based on the linkage homophily principle;  
  - proposed an **iterative framework** to predict multiple types of links collectively.

- Performances of link prediction can be improved by collectively predicting multiple types of links in HINs.