Link Prediction using Machine Learning Algorithms

Zhang Meng

School of Computer Science, Hubei University of Technology

Abstract - *With the advent on Internet, research on social network has improved in a rapid pace. In the context of Social Network Analysis (SNA), link prediction has become an important research direction. In this paper, we use supervised learning along with social network metrics to improve the accuracy of the link prediction task. We used multiple machine learning algorithms that are commonly used for prediction task. We also improved the classifiers features by extracting and adding multiple SNA metrics such as centrality metrics. We analyzed our method on a dataset of 2000 authors that occasionally collaborated with each other to write papers. Our analysis showed that enhancing the feature list by SNA metrics increased the accuracy of the prediction task.*

Key Words: link prediction, machine learning, social network analysis, SNA, artificial neural network, co-authorship

1. INTRODUCTION

With the advent of Web 2.0 and social web, complex social networks have emerged. Social networks are defined as structures that nodes represent entities (e.g. people) and edges represent any relationship between nodes (e.g. interaction, friendship, collaboration) [1]. Examples of social networks includes friendship networks in social media like Instagram, authorship network between authors of the papers [2]. The research on social network has increased in the past decade [1, 3, 4]. One of the branches of social network research is focused on the evolution of the existing networks and predicting the links between nodes that may interact in the future [5].

In this paper, we evaluate the performance of different machine learning algorithms in link prediction task. In order to improve the accuracy of the prediction task, we employed many social network analysis metrics, such as closeness, betweenness. To predict the links between entities, we applied multiple machine learning algorithms that are used in many successful studies [6-9].

The remainder of this paper organized as follows. Section 2 reviews a brief background of link prediction task and social network analysis. Section 3 describes the collected dataset and feature extraction process. Section 4 discusses the results of using different machine learning algorithms. Finally, Section 5 concludes the paper.

2. BACKGROUND

In this section we review the backgrounds of link prediction and social network analysis.

2.1 Link Prediction

Link prediction is defined as the problem of predicting new or deleted links between nodes of a social network in the future [10]. This problem also can be defined as the problem of estimating the likelihood of the existence of a link between two nodes, based on the observed links and characteristics of the network [5]. Fig 1 shows an example of predicting a link between nodes 1 and 3.

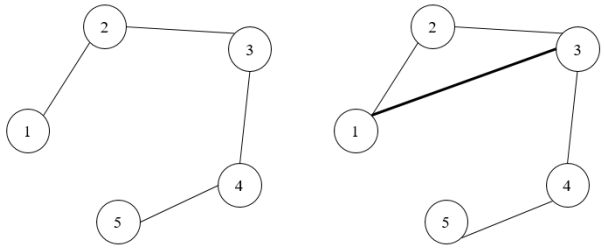

Fig -1: Link prediction between nodes 1 and 3

Research on link prediction has been increased over the past decade [11-14]. Link prediction has a vast domain of applications in different fields. Recommender systems can apply link prediction techniques to identify most similar items [15]. Robust online information extraction methods in recommender systems [9] creates a great demand for link prediction algorithms. Social networks can use link prediction techniques to expose users to most similar users to expand the network. Another application of link prediction is for bot detection using transformers model in social media [16]. Research on Biology also uses link prediction techniques to identify possible interactions between proteins [17].

2.2 Social Network Analysis

Social networks have become an internal parts in today's lives. The extensive usage of social network led to increase in the value of the data available on the network. The research on Social Network Analysis (SNA) has been increased over the past decades. Formally a social network is represented as a graph $G = <V,E>$ in which each edge $e=<u,v> \in E$ represents

Table -1: Original Dataset

author1	author2	pages	year	volume	journal	Ee
Zong Ming	Bin Cai	79	2011	12	BMC Bioinformatics	https://doi.org/10.1186/1471-2105-12-79
Zong Ming	C. Neal Stewart Jr.	3	2009	10	BMC Bioinformatics	https://doi.org/10.1186/1471-2105-10-S11-S3
Zong Ming	Crissa Doeppke	3	2009	10	BMC Bioinformatics	https://doi.org/10.1186/1471-2105-10-S11-S3

a relationship (e.g. friendship, authorship, and etc.) between two nodes $u,v \in V$.

Social network analysis has been widely used in many studies. Social network analysis can be used for modeling the dynamics of complex networks [4]. One of the applications of SNA is in improving recommender systems. Ebrahimi et al. [1] used SNA techniques to improve the similarities between nodes in recommender systems. Oghaz et al. [18] used SNA for topic mining considering temporal-causal relationships among events. Another application of the SNA is for improving link prediction task [2].

Applications of SNA used SNA metrics to improve the desired tasks. Studies show that SNA metrics can be used to represent the influence of users on the network [3]. One of the main categories of SNA metrics that used in many studies is centrality metrics, like closeness, degree, and eigenvector centralities [19].

3. DATA COLLECTION

In this section, we describe the process for collecting and processing the data set used in our experiments.

3.1 Dataset

For our analysis we used DBLP dataset. The DBLP Computer Science Bibliography evolved from an early small experimental Web server to a popular service for the computer science community [20].

The DBLP dataset provides an easy way to derive graphs like coauthor graph, which is an example of a social network. For this purpose, we use the data set is available from http://dblp.uni-trier.de/xml/. The file dblp.xml contains all bibliographic records which make DBLP. It is accompanied by the data type definition file dblp.dtd. We need this auxiliary file to read the XMLfile with a standard parser. dblp.xml has a simple layout:

```
<?xml version="1.0" encoding="utf-8"?>
<dblp>
<article mdate="2017-05-28" key="journals/acta/Saxena96">
<author>Sanjeev Saxena</author>
<title>Parallel Integer Sorting and Simulation Amongst CRCW Models.</title>
 <pages>607-619</pages>
<year>1996</year>
<volume>33</volume>
<journal>Acta Inf.</journal>
<number>7</number>
<url>db/journals/acta/acta33.html#Saxena96</url>
<ee>https://doi.org/10.1007/BF03036466</ee>
```

3.2 Data Processing

From the raw data set, we extract the pairs of authors and other attributes like number of pages, the year of publication, the journal, the volume, and the doi. As represented in the Table 1, there is no attribute that can help for predicting co-authoring. To this end, we processed the data to extract some attributes. First, we counted the number of publications of each author and the number of common publications between two authors. Table 1 and Table 2, shows the extracted data. As the DBLP dataset was very large, we decided to select only 20000 co-authorship pairs randomly.

Table -2: Number of Publications by each Author

Author	No of Publications
Azriel Rosenfeld	42
Ness B. Shroff	36
R. Srikant	31
Donald F. Towsley	28
Ian F. Akyildiz	26
…	…

Table -3: Number of Common Publications between Two Authors

Author1	Author2	No of common Publications
Anna Spagnolli	Luciano Gamberini	115
Anna Spagnolli	Giuseppe Riva	66
Giuseppe Riva	Luciano Gamberini	60
Achim Jung	Anna Spagnolli	33
Anna Spagnolli	John A. Waterworth	33
John A. Waterworth	Luciano Gamberini	30
Achim Jung	Luciano Gamberini	30
…..	…..	….

author1	Degree	Closeness Centrality	Betweenness Centrality	Page Rank	author2	Degree2	Closeness Centrality2	Betweenness Centrality2	Page Rank2	No of Common Publications
Anna Spagnolli	0.40708	0.149391	5.70E-05	0.00042	Luciano Gamberini	0.54867	0.149416	0.000721	0.000424	115
Anna Spagnolli	0.40708	0.149391	5.70E-05	0.00042	Giuseppe Riva	0.40708	0.149391	0.000194	0.000259	66
Giuseppe Riva	0.40708	0.149391	0.000194	0.000259	Luciano Gamberini	0.54867	0.149416	0.000721	0.000424	60
……	……	……	……	……	……	……	……	……	……	……

Table -4: Final Format of the dataset

Since the dataset does not provide any attribute, except the number of publications, we used Social Network Analysis to add some features to the dataset. In the constructed SNA, nodes are the authors and each link represents that two authors have a common publication. Considering the 20000 co-authorship links as the links of the network, Fig 2 shows an overall image of the constructed network. This network has 15908 nodes (authors) and 20000 edges (co-authorship relation) and nodes with darker colors have larger degrees.

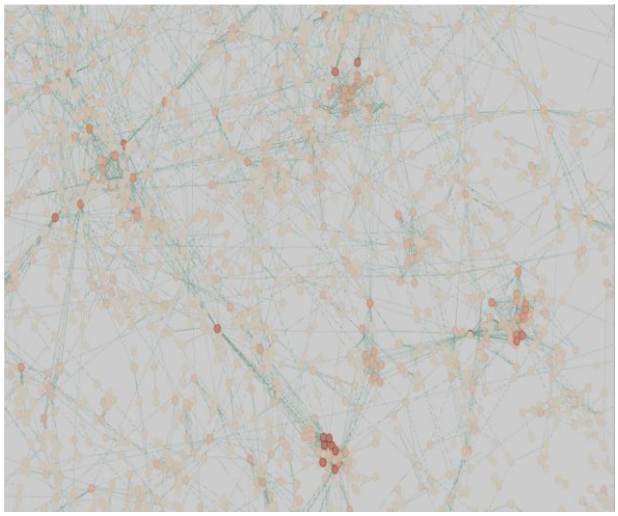

Fig -2: The Constructed Social Network

After analyzing the network, for each node, we calculated its centrality metrics like degree, closeness, and betweenness. In addition, we added the PageRank of each node to the data. Consider that in the created network, the degree of each node represents the number of publications of that author, so we did not consider this feature twice. Therefore, the attributes of our dataset are [Degree1, Closeness1, Betweenness1, PageRank1, Degree2, Closeness2, Betweenness2, PageRank2, Class]. The final format of dataset are shown in Table 4.

To define the class, first we use the number of common publications for the linear regression model. But for the classification algorithms the class cannot be numerical. We mapped the number of common publications to a categorical label. It has been indicated in Table 5.

Table -5: Definition of Classes

Number of common publications	Class
> 8	High
>1	Medium
=1	Low

4. ANALYSIS

For classification task, we used the most common algorithms used for machine learning tasks [8, 21]. We used linear regression, artificial neural network, k nearest neighbors, and naïve bayes. Here in this section we implement these algorithms and report the results.

4.1 Linear Regression

As we mentioned earlier, we did not map the number of common publications for this algorithm. So here we want to predict the number of common publications between two authors based on their attributes. First, we only consider two attributes (equation 1): the degree of author1 (the number of author1's publications) and the degree of author2 (the number of author2's publications).

# of common publications = 1.31 + 1.5 * Degree1 + 2.06 * Degree2 (Equation 1)

Table -6: Classes Calculated by Linear Regression Considering Degree 1 and Degree 2

Degree1	Degree2	# of common publications
0.40708	0.548673	115
0.40708	0.40708	66
0.40708	0.548673	60
0.380531	0.40708	33
0.40708	0.380531	33

Table -7: Classes Calculated by Linear Regression Considering all Features

Degree	Closeness Centrality	Betweenness Centrality	Page Ranks	Degree2	Closeness Centrality2	Betweenness Bentrality2	Page Ranks2	# of common publications
0.40708	0.149391	5.70E-05	0.00042	0.548673	0.149416	0.000721	0.000424	115
0.40708	0.149391	5.70E-05	0.00042	0.40708	0.149391	0.000194	0.000259	66
0.40708	0.149391	0.000194	0.000259	0.548673	0.149416	0.000721	0.000424	60
0.38053	0.149424	0.000802	0.000143	0.40708	0.149391	5.70E-05	0.00042	33
0.40708	0.149391	5.70E-05	0.00042	0.380531	0.149388	5.60E-05	0.000133	33
…	…	…	…	…	…	…	…	…

The best result of considering only these two attributes gained according to the Equation 1. The resulted mean absolute error was 0.76. The results have been indicated in Table 6. Then we added other SNA features to the dataset. This change leads to change of mean absolute error to 0.62. The new classes are shown in Table 7.

4.2 Artificial Neural Network

ANN is another widely-used machine learning algorithm that is used in many studies. Now we have the categorical values (high, medium, low) of classes, and we will follow some classification algorithms. The constructed ANN has been indicated in Fig 3.

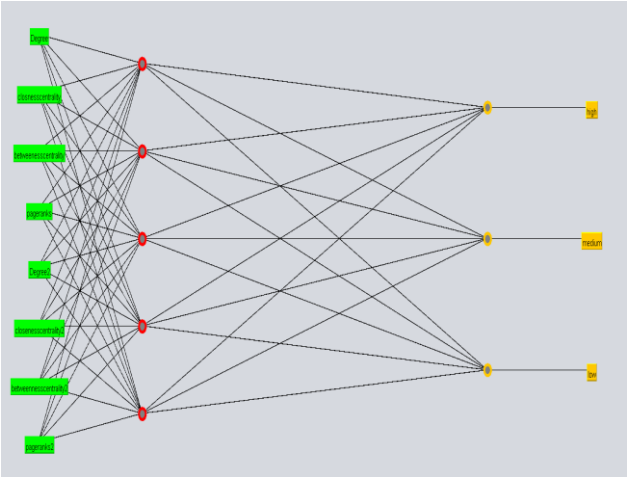

Fig -3: The Constructed ANN

For all the algorithms, we will run the algorithm for two conditions:

- Considering all SNA metrics (Table 8)
- Considering only two attributes: Degree of authors (or Number of publications) (Table 9)

Table -8: Accuracy obtained Using ANN Considering Degrees

	Precision	Recall	F-Measure
High	0.45	0.63	0.62
Medium	0.37	0.11	0.17
Low	0.87	1	0.93
Average	0.69	0.58	0.57

Table -9: Accuracy obtained Using ANN Considering all Features

	Precision	Recall	F-Measure
High	0.892	0.565	0.692
Medium	0.605	0.313	0.413
Low	0.913	0.976	0.944
Average	0.877	0.895	0.897

By changing the value of learning rate and momentum, the accuracy measurements are changed as it has been shown in Table 10. The best result achieved when we set learning rate to 0.01, and momentum to 0.1.

Table -10: Accuracy Measurements Obtained Using ANN by Changing Learning Rate and Momentum

a= 0.001 & m=0.2				a=0.01 & m=0.2			
	Precision	**Recall**	**F-Measure**		**Precision**	**Recall**	**F-Measure**
High	0.387	0.634	0.483	**High**	0.835	0.596	0.696
Medium	0.432	0.123	0.191	**Medium**	0.597	0.305	0.404
low	0.893	0.987	0.937	**Low**	0.913	0.975	0.943
average	0.57	0.877	0.522	**average**	0.875	0.893	0.878
a=0.1 & m=0.01				a=0.1 & m=0.2			
	Precision	**Recall**	**F-Measure**		**Precision**	**Recall**	**F-Measure**
High	0.882	0.559	0.684	**High**	0.883	0.565	0.689
Medium	0.596	0.295	0.395	**Medium**	0.602	0.301	0.401
low	0.911	0.976	0.943	**Low**	0.912	0.976	0.943
average	0.874	0.893	0.876	**average**	0.875	0.894	0.877
a=0.1 & m=0.4				a=0.3 & m=0.2			
	Precision	**Recall**	**F-Measure**		**Precision**	**Recall**	**F-Measure**
High	0.883	0.565	0.689	**High**	0.451	0328	0.37
Medium	0.606	0.298	0.400	**Medium**	0.762	0.532	0.623
Low	0.912	0.977	0.943	**Low**	0.874	1	0.933
average	0.876	0.894	0.877	**Average**	0.695	0.874	0.642

Table -11: Accuracy Measurements Obtained Using KNN by Changing K Considering Only Degrees

K=1				K=3			
	Precision	**Recall**	**F-Measure**		**Precision**	**Recall**	**F-Measure**
High	0.467	0.484	0.476	**High**	0.455	0.466	0.460
Medium	0.328	0.167	0.221	**Medium**	0.351	0.126	0.186
Low	0.895	0.954	0.924	**Low**	0.892	0.968	0.928
average	0.825	0.858	0.838	**Average**	0.824	0.865	0.837
K=6				K=10			
	Precision	**Recall**	**F-Measure**		**Precision**	**Recall**	**F-Measure**
High	0.537	0.410	0.465	**High**	0.505	0.342	0.407
Medium	0.344	0.095	0.149	**Medium**	0.372	0.054	0.094
Low	0.888	0.976	0.930	**Low**	0.884	0.988	0.933
average	0.821	0.868	0.834	**Average**	0.820	0.873	0.830

Table -12: Accuracy Measurements Obtained Using KNN by Changing K Considering All Features

K=1				K=3			
	Precision	**Recall**	**F-Measure**		**Precision**	**Recall**	**F-Measure**
High	0.772	0.714	0.742	**High**	0.699	0.720	0.709
Medium	0.532	0.536	0.534	**Medium**	0.537	0.443	0.485
Low	0.938	0.938	0.938	**low**	0.928	0.949	0.938
average	0.889	0.889	0.889	**average**	0.880	0.888	0.883
K=6				K=10			
	Precision	**Recall**	**F-Measure**		**Precision**	**Recall**	**F-Measure**
High	0.829	0.634	0.718	**High**	0.926	0.547	0.688
Medium	0.526	0.452	0.486	**Medium**	0.560	0.373	0.448
Low	0.928	0.947	0.937	**low**	0.919	0.963	0.941
average	0.880	0.887	0.883	**average**	0.877	0.891	0.881

4.3 K Nearest Neighbors

By considering different values of K, accuracy measurements will be changed as it has been indicated in Table 11 and Table 12.

4.4 Naïve Bayes

We have used this algorithm to analyze data in two ways: 1. Considering only degrees (Table 13) and 2. Considering all features (Table 14).

Table -13: Accuracy obtained Using Naïve Bayesian Considering Degrees

	Precision	**Recall**	**F-Measure**
High	0.079	0.199	0.113
Medium	0.238	0.102	0.143
Low	0.890	0.946	0.917
Average	0.807	0.840	0.819

Table -14: Accuracy obtained Using Naïve Bayesian Considering all Features

	Precision	**Recall**	**F-Measure**
High	0.148	0.720	0.245
Medium	0.248	0.208	0.227
Low	0.906	0.893	0.900
Average	0.823	0.812	0.815

5. DISCUSSION & CONCLUSION

Comparing different algorithm for classification of dataset, results shows that KNN with k=6 when all features are considered, is the best one for classifying. After KNN, ANN worked better, while the results of Naïve Bayesian were not promising.

For all of the applied algorithms adding SNA metrics to the prediction model, increased the accuracy. But for Naïve Bayesian, adding SNA metrics as features did not change the accuracy much. The reason can be related to the small number of features that we had. Also, most of the used features were correlated.

Another important consideration is that, the average number of common publications between pairs of authors was 1.81 which is near to class “low”. Therefore we cannot conclude that our classifiers worked very well to classify instances in these category. Basically, when we wanted to compare algorithms with each other, we mostly looked at the “medium” and “high” classes, since their accuracy were more representative of the overall accuracy.

REFERENCES

[1] F. Ebrahimi and S. A. H. Golpayegani, "Personalized recommender system based on social relations," in 2016 24th Iranian Conference on Electrical Engineering (ICEE), 2016: IEEE, pp. 218-223.

[2] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," Journal of the American society for information science and technology, vol. 58, no. 7, pp. 1019-1031, 2007.

[3] I. Garibay, A. V. Mantzaris, A. Rajabi, and C. E. Taylor, "Polarization in social media assists influencers to become more influential: analysis and two inoculation strategies," Scientific reports, vol. 9, no. 1, pp. 1-9, 2019.

[4] I. Garibay et al., "Deep Agent: Studying the Dynamics of Information Spread and Evolution in Social Networks," arXiv preprint arXiv:2003.11611, 2020.
[5] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," Physica A: statistical mechanics and its applications, vol. 390, no. 6, pp. 1150-1170, 2011.
[6] M. Rezaei, E. Ravanbakhsh, E. Namjoo, and M. Haghighat, "Assessing the Effect of Image Quality on SSD and Faster R-CNN Networks for Face Detection," in 2019 27th Iranian Conference on Electrical Engineering (ICEE), 2019: IEEE, pp. 1589-1594.
[7] R. E. Meymand, A. Soleymani, and N. Granpayeh, "All-optical AND, OR, and XOR logic gates based on coherent perfect absorption in graphene-based metasurface at terahertz region," Optics Communications, vol. 458, p. 124772, 2020.
[8] F. Jafariakinabad, S. Tarnpradab, and K. A. Hua, "Syntactic recurrent neural network for authorship attribution," arXiv preprint arXiv:1902.09723, 2019.
[9] M. Heidari and S. Rafatirad, "Using Transfer Learning Approach to Implement Convolutional Neural Network to Recommend Airline Tickets by Using Online Reviews," presented at the International Workshop on Semantic and Social Media Adaptation and Personalization, 2020.
[10] P. Wang, B. Xu, Y. Wu, and X. Zhou, "Link prediction in social networks: the state-of-the-art," Science China Information Sciences, vol. 58, no. 1, pp. 1-38, 2015.
[11] A. Soleymani, R. E. Meymand, and N. Granpayeh, "Broadband near-perfect terahertz absorber in single-layered and non-structured graphene loaded with dielectrics," Applied Optics, vol. 59, no. 9, pp. 2839-2848, 2020.
[12] M. Pavlov and R. Ichise, "Finding experts by link prediction in co-authorship networks," FEWS, vol. 290, pp. 42-55, 2007.
[13] Z. Liu, Q.-M. Zhang, L. Lü, and T. Zhou, "Link prediction in complex networks: A local naïve Bayes model," EPL (Europhysics Letters), vol. 96, no. 4, p. 48007, 2011.
[14] M. Sachan and R. Ichise, "Using semantic information to improve link prediction results in network datasets," International Journal of Engineering and Technology, vol. 2, no. 4, p. 334, 2010.
[15] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative filtering recommender systems," in The adaptive web: Springer, 2007, pp. 291-324.
[16] M. Heidari and J. H. J. Jones, "Using BERT to Extract Topic-Independent Sentiment Features for Social Media Bot Detection," presented at the Annual Ubiquitous Computing, Electronics & Mobile Communication Conference, 2020.
[17] Y. Qi, Z. Bar-Joseph, and J. Klein-Seetharaman, "Evaluation of different biological data and computational classification methods for use in protein interaction prediction," Proteins: Structure, Function, and Bioinformatics, vol. 63, no. 3, pp. 490-500, 2006.
[18] T. A. Oghaz, E. C. Mutlu, J. Jasser, N. Yousefi, and I. Garibay, "Probabilistic Model of Narratives Over Topical Trends in Social Media: A Discrete Time Model," arXiv preprint arXiv:2004.06793, 2020.
[19] A. Rajabi, C. Gunaratne, A. V. Mantzaris, and I. Garibay, "On Countering Disinformation with Caution: Effective Inoculation Strategies and Others that Backfire into Community Hyper-Polarization," in International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, 2020: Springer, pp. 130-139.
[20] M. Ley, "DBLP: some lessons learned," Proceedings of the VLDB Endowment, vol. 2, no. 2, pp. 1493-1500, 2009.
[21] S. Goudarzvand, J. S. Sauver, M. M. Mielke, P. Y. Takahashi, Y. Lee, and S. Sohn, "Early temporal characteristics of elderly patient cognitive impairment in electronic health records," BMC medical informatics and decision making, vol. 19, no. 4, p. 149, 2019.