DA516 Social Network Analysis Term Project Proposal

Spring 2019-2020

M.Sc. in Data Analytics-Sabanci University

A Basic Recommendation System Based on Link Prediction (Retail Example)

Contents

1. Team Members

Sinan Türkmen (sinanturkmen@sabanciuniv.edu), Güzin Erdem (guzinerdem@sabanciuniv.edu)

2. Data

a. Data Description

Amazon-Rating Recommendation Network

The ratings represent edge weights and there is additional column for the edge timestamp.

- Category: Sparse Networks, Recommendation Networks
- Short: Users-rate-products
- Vertex type: User, product
- 4. Edge type: Rating, edge attribute
- Format: Bipartite
- Edge weights: Weighted

b. Data Reference

@inproceedings{nr,

title={The Network Data Repository with Interactive Graph Analytics and Visualization},

author={Ryan A. Rossi and Nesreen K. Ahmed},

booktitle={AAAI},

url={http://networkrepository.com},

year={2015}

}

3. Introduction and Problem Definition

The objective of link prediction is to identify pairs of nodes that will either form a link or not in the future. Link prediction is used in tons of real-world application. In our project, we try to predict customer who are likely to buy products on online market and right after to make a product recommendation.

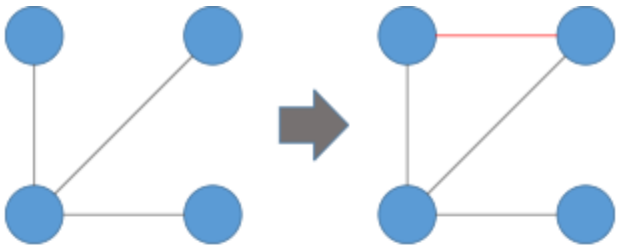
In this project, similarity measures and supervised learning algorithms are compared inside a real network.

4. Literature Review

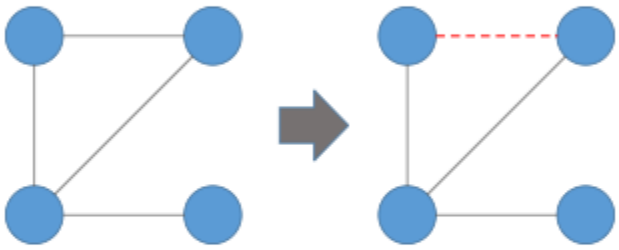
Link prediction problem has been extensively studied by members of the complex network community. Liben-Nowell and Kleinberg have formalized the link prediction problem in the following way. Let G(V,L) be a network within the time period of G(t,t1) where V represents the set of nodes and L represents the set of links. For the next time period G(t1,t2), the network might change. The link prediction focuses on how to predict the evolution of links, that is, how L(t,t1) will differ from L(t1,t2).

Researchers with background in physics and mathematics usually deal with the problem by focusing on the topology information of the networks. Researchers with machine learning and data mining background favour to solve the problem with considering the nodes' attribute information. There are three types of link prediction problems as shown in Figure 1: we can consider (i) only adding links to the existing network, (ii) only removing links from the existing structure, and (iii) both, adding and removing links at the same time.

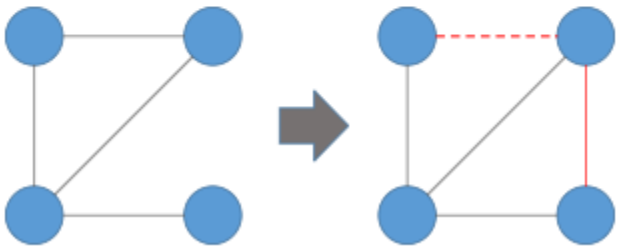
In this research, we will only focus on the first type of link prediction problem which only aims at predicting the appearance of links.[i]



(a) Adding links



(b) Removing links



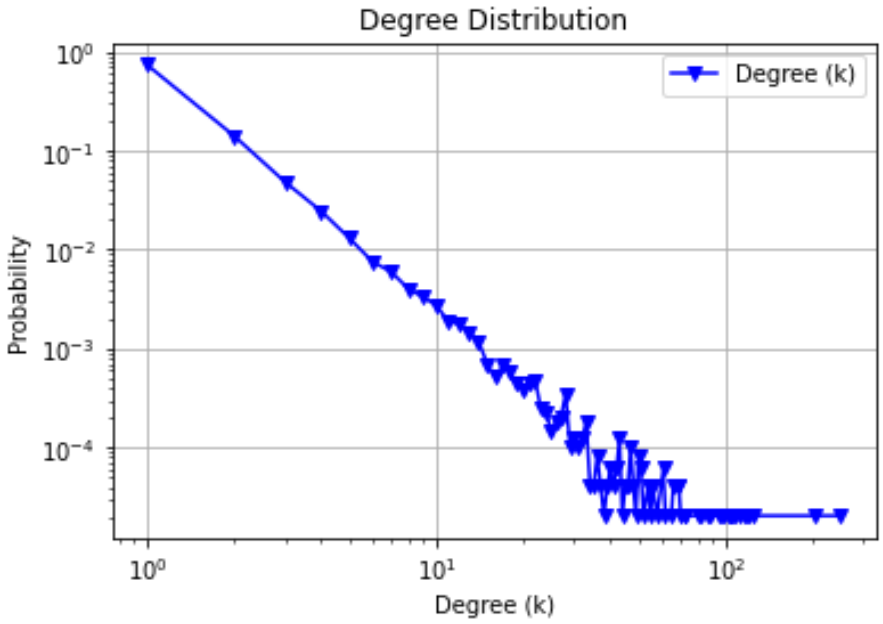
(c) Adding and removing links

5. Exploratory Analysis and Data Preprocessing

The number of edges is 5.8 Million and the number of nodes is 2.1 Million. Additionally, the average degree of our network is 5 while the clustering coefficient is almost 0.0004 which is quite small as we expect. Since the data is too large to conduct some algorithms on it, we firstly focus on users giving rating 4 and 5 to products. As a result, the number of nodes and edges become 1.93 Million and 4.45 Million, respectively. Then, we select 10 % random sample based on products. After random sampling, we get 46,029 edges and 193,942 nodes in the sample.

Density is the number of connections a node has, divided by the total possible connections a participant could have. Density of the network is nearly zero which is too low, indeed. The case is normal for our system since users do not have to vote a product each time. Thus, our network is a sparse network, and low density is expected result.

Further, degree distribution covers a small amount of information about a network. However, information gives some important clues about the structure of a network.



As we can see from the distribution plot, there exist the nodes with very high degree. This implies the network has power law degree distribution. It is long tail distribution. Large-degree nodes are often referred to as hubs so there are probably hubs having connections to many others in the system. The average degree is nearly 1.8 which is again low. In a real-world network, most nodes have a relatively small degree. Therefore, we already expected a low average degree in our network.

Furthermore, one of the features of a node is clustering coefficient. It gives information about the neighborhood of a node. It implies how well connected the neighbors of a node. The

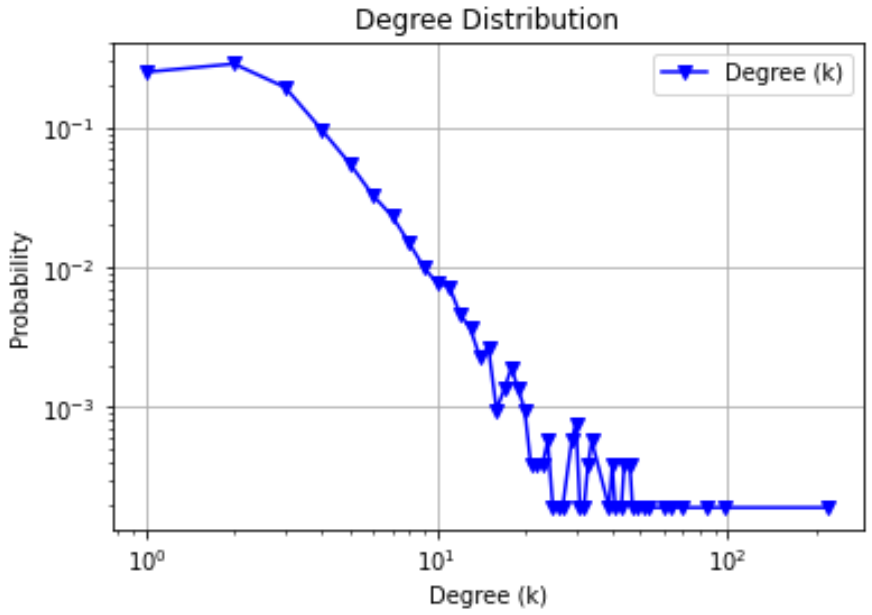
clustering coefficient value is almost zero for our network. This tells that the possibility to have connections in the neighborhood is almost zero.

Moreover, our network is not fully connected. However, it includes connected components. The number of connected components of the system is 9,317. When we firstly looked at the biggest connected component, its size is 27,510. Namely, it includes 27,510 nodes and 32,781 edges. it is 54 % of the whole network.

Then, we focused on the nodes whose degree is greater than 3. Thus, we removed the nodes having degree less than and equal to 3. However, it would be still computationally expensive we again took the giant component from the remaining sample network. We got eventually 5,225 nodes and 8,972 edges in our component which is our full dataset to make prediction.

After all, shortest path is a measure for efficiency of information on a network. The average number of steps along the shortest paths for all possible pairs of nodes in our final component is 5.83. This implies a bit long path length between all pairs of nodes. Thus, this provides less efficient transmission in the largest connected component of our network. Further, its diameter is 19 which simply implies a long longest of all geodesics in the giant component. This gives us an idea how far a product from another.

Besides, it has 3.43 average degree that is larger than the system and the degree distribution of final sample is below:



The plot again points out largest component also has a power law distribution. In addition to this, there are some nodes with high degrees, which refers to possibility of having hubs.

The schema below gives the graph sizes and their preprocessed operations.

Name	Operations done	# of nodes	# of edges
G-Full Dataset	-	2.1 M	5.38 M
G	Ratings of 4 and 5	1.93 M	4.45 M
G1	10% sampling	193,756	46,029
G2	Giant component	190,198	44,206
G3	Degree greater than or equal to 3	5,225	8,972
Gtrain		***5,225***	***7,178***
Gtest			***1,794***

6. Methodology

a. Similarity Measures

Firstly, we implement most frequent category in link prediction which is similarity – based methods. The assumption in similarity – based methods is that two nodes have more intent to interact if they are similar to each other based on a defined similarity function. Therefore, we implement Degree Centrality, SimRank, Preferential Attachment, Adamic – Adar and Jaccard similarity functions to measure local similarities.

- **SimRank algorithm** measures the structural similarities of entities. Its underlying assumption is that entities are similar if they are connected to similar objects.
- **Preferential Attachment** is a measure to compute closeness of nodes based on their shared neighbors. Its intuition hinges upon network entities prefer to make a connection to the more popular existing ones.
- **Adamic – Adar similarity algorithm** measures the amount of shared links between two entities. It is actually called as inverse logarithmic degree centrality of shared neighbors by two nodes
- **Jaccard similarity index** compares two nodes to see shared neighbors.

After experiencing these similarity functions, we initially tried different thresholds on Similarity Rank scores to make prediction.

Since, some rankings can give unusual thresholds they are normalized by dividing them to maximum similarity score value.

b. Supervised Binary Classification

As a second method, we try supervised learning technique with different classifiers on the graph.

We split it into 80% train and 20 % test based on edges existing in the network for making prediction. Then, we append a similar size of negative set that refers to pairs of nodes not

connected by edges. The rest of negative edges become the part of test set. Eventually, we get 5,225 nodes and 7,178 edges in the train set and 1,794 edges in the test set.

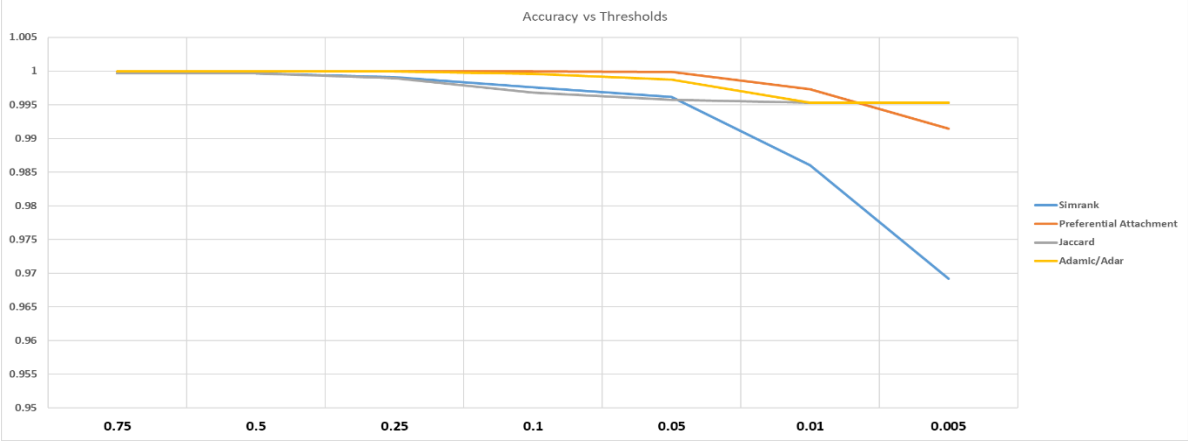
Additionally, our train matrix includes the similarity metrics which we extract from the network. These metrics are ***SimRank, Degree Centrality, Jaccard Similarity index, Preferential Attachment index and Adamic – Adar Similarity score***. Our dependent variable includes positive or negative labels implying that edges are in train or not.

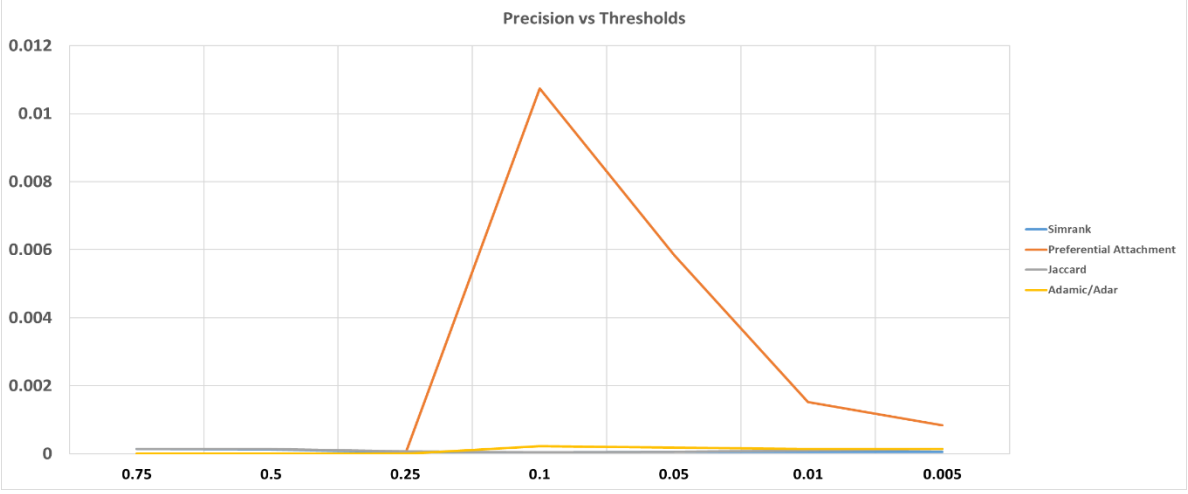
Since similarity scores are in different scales, we ***do standard scaling to train dataset***. We then make train XGBoost Classifier, Random Forest Classifier, Decision Tree Classifier and Stochastic Gradient Descent Classifier.

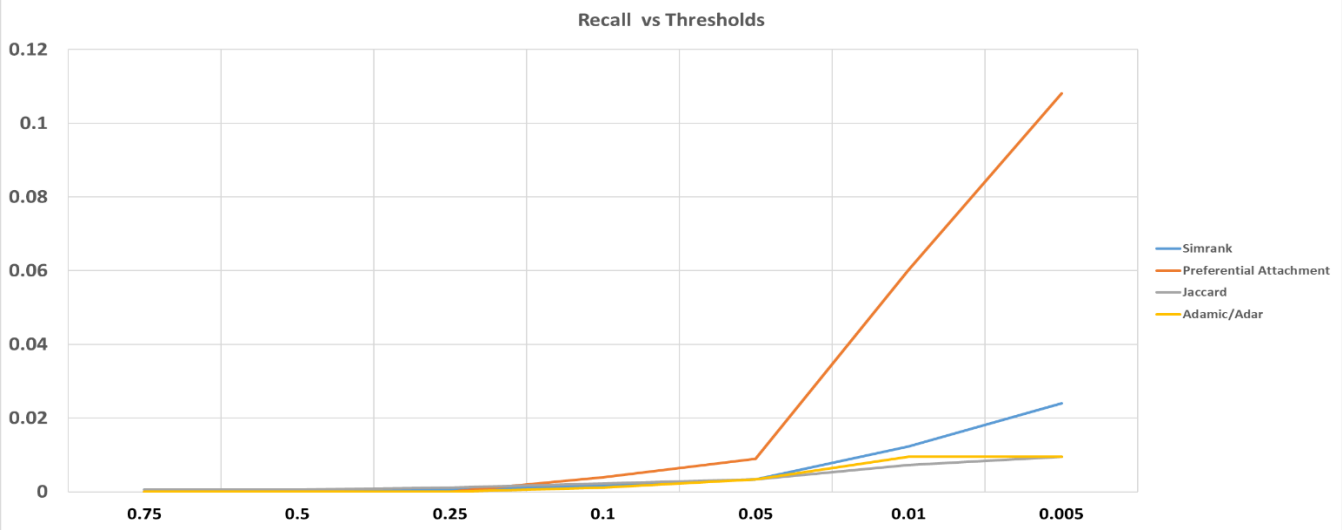
7. Results

a. Similarity Measures

After working on train and test data, different, precision/recall and accuracy values are seen. The graphs below gives the related values.

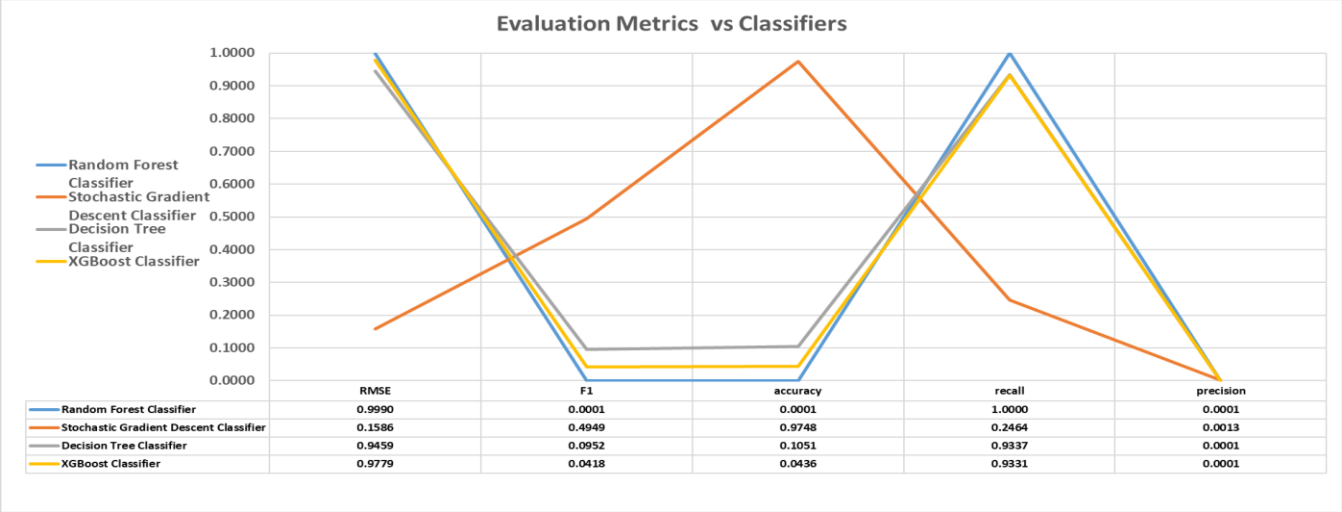






b. Supervised Binary Classification

Worked classifiers and their performances can be seen in the graph below.



	RMSE	F1	accuracy	recall	precision
Random Forest Classifier	0.9990	0.0001	0.0001	1.0000	0.0001
Stochastic Gradient Descent Classifier	0.1586	0.4949	0.9748	0.2464	0.0013
Decision Tree Classifier	0.9459	0.0952	0.1051	0.9337	0.0001
XGBoost Classifier	0.9779	0.0418	0.0436	0.9331	0.0001

8. Evaluation and Comments

As it can be seen from the results, supervised binary classification works much better than similarity measures. Even though, supervised learning composes of those similarity measures, its scores especially trade-off between precision and recall is better.

However, in even supervised learning precision scores are so small that appliers that worry about false positive rates should handle with this problem. As we can see from our researches, recall rate is much more important in sectors like retail, since a person with an opportunity to buy is an advantage to increase turnover values.

9. References

[i] Link Prediction Methods and Their Accuracy for Different Social Networks and Network Metrics, Fei Gao,1 Katarzyna Musial ,1 Colin Cooper,1 and Sophia Tsoka1