CENTRAL UNIVERSITY OF RAJASTHAN

Link Prediction in Complex Network

PRESENTED BY:

CHARUL MERTIA
2022MSCS006
DEPARTMENT OF COMPUTER SCIENCE

UNDER THE SUPERVISION OF:

Dr. Abhay Kumar Rai
Assistant Professor
DEPARTMENT OF COMPUTER SCIENCE

CONTENTS

- Datasets Used
- Understanding the Data
- Dataset Preparation for Model Building
- Feature Extraction
- Building Link Prediction Model
- Logistic Regression
- LightGBM
- Accuracy Scores
- AUC-ROC Score
- References

Datasets Used

- Fb-Pages-Food : the nodes are Facebook pages of popular food joints and well-renowned chefs from across the globe. If any two pages (nodes) like each other, then there is an edge (link) between them. [6.1]
- Fb-Pages-Tvshow : Data collected about Facebook pages. These datasets represent blue verified Facebook page networks of different categories. Nodes represent the pages and edges are mutual likes among them. [6.2]
- Arenas-Email : This is the email communication network of a University in Spain. Nodes are users and each edge represents that at least one email was sent. The direction of emails or the number of emails are not stored. [6.3]

Understanding the Data

1. Import all the necessary libraries and modules: pandas, numpy, matplotlib, sklearn, networkx.

2. Load the datasets. Number of Links:

    1. Fb-Pages-Food : 2102

    2. Fb-Pages-Tvshow : 17262

    3. Arenas-Email : 5451

3. Create a dataframe of all the nodes.

Dataset Preparation for Model Building

1. Retrieve Unconnected Node Pairs – Negative Samples
    - Create an adjacency matrix.
    - Traverse adjacency matrix to get the count of unconnected node pairs.
        - Fb-Pages-Food : 19018
        - Fb-Pages-Tvshow : 96887
        - Arenas-Email : 59345
    - Add all the above pairs in a dataframe.

Dataset Preparation for Model Building

2. Remove Links from Connected Node Pairs – Positive Samples
    - Drop edges in a way that all the nodes of the graph should remain connected.
        - Fb-Pages-Food : 1483
        - Fb-Pages-Tvshow : 13371
        - Arenas-Email : 4319
3. Data for Model Training
    - Append the removable edges to the dataframe of unconnected node pairs.
    - Distribution of values of the target variable:
        - Fb-Pages-Food : 0 : 1010 1 : 990
        - Fb-Pages-Tvshow : 0 : 6868 1 : 9132
        - Arenas-Email : 0 : 3852 1 : 4148

Feature Extraction

- Algorithms to extract node features:

1. Adamic-Adar Index
2. Preferential Attachment
3. The Katz Index (KI)
4. SimRank (SR)
5. The Local Path Index (LPI)
6. PropFlow Predictor (PFP).

Building Link Prediction Model

- Split the data into two parts – training set and testing set.
- Use Machine Learning algorithm on the model and check the performance.
    1. Logistic Regression
    2. LightGBM

Logistic Regression [3]

- It is a linear model used for binary classification problems, where the target variable has two possible outcomes.
- The primary goal of logistic regression is to model the relationship between a dependent binary variable and one or more independent variables (features).
- Unlike linear regression, which predicts continuous outcomes, logistic regression predicts the probability of a binary outcome (usually coded as 0 or 1).

LightGBM [4]

- **LightGBM** is a gradient boosting framework that uses tree based learning algorithms.
- Advantages:
  - Faster training speed and higher efficiency.
  - Better accuracy.
  - Support of parallel, distributed, and GPU learning.
  - capable of handling large-scale data.
- Parameters:
  - 'objective': ‘binary’
  - early_stopping(stopping_rounds= 20)

Accuracy Scores

Algorithms	Models	Fb-Pages-Food	Fb-Pages-Tvshow	Arenas-Email
Adamic-Adar Index	Logistic Regression	0.5700	0.5663	0.5113
	LightGBM	0.5700	0.5663	0.5146
Preferential Attachment	Logistic Regression	0.6000	0.5227	0.5438
	LightGBM	0.6200	0.5223	0.5146
The Katz Index	Logistic Regression	0.5417	0.5221	0.5221
	LightGBM	0.6067	0.5529	0.5529
SimRank	Logistic Regression	0.5700	0.5663	0.5113
	LightGBM	0.5700	0.5663	0.5146
The Local Path Index	Logistic Regression	0.5683	0.5658	0.5113
	LightGBM	0.5683	0.5663	0.5146
PropFlow Predictor	Logistic Regression	0.5700	0.5663	0.5113
	LightGBM	0.5700	0.5663	0.5146

AUC-ROC Scores

Algorithms	Models	Fb-Pages-Food	Fb-Pages-Tvshow	Arenas-Email
Adamic-Adar Index	Logistic Regression	0.577155	0.616831	0.524910
	LightGBM	0.577205	0.616697	0.524984
Preferential Attachment	Logistic Regression	0.671009	0.553403	0.568931
	LightGBM	0.666880	0.576009	0.571605
The Katz Index	Logistic Regression	0.476043	0.512508	0.512508
	LightGBM	0.692842	0.583940	0.583940
SimRank	Logistic Regression	0.576999	0.616886	0.524955
	LightGBM	0.576994	0.616885	0.524955
The Local Path Index	Logistic Regression	0.575354	0.616517	0.524955
	LightGBM	0.575354	0.616517	0.524955
PropFlow Predictor	Logistic Regression	0.577155	0.616831	0.524910
	LightGBM	0.577183	0.616697	0.524984

References

1. Martínez, Víctor, Fernando Berzal, and Juan-Carlos Cubero. "A survey of link prediction in complex networks." *ACM computing surveys (CSUR)* 49.4 (2016): 1-33.

2. https://www.analyticsvidhya.com/blog/2020/01/link-prediction-how-to-predict-your-future-connections-on-facebook/

3. https://scikit-learn.org/stable/index.html

4. https://lightgbm.readthedocs.io/en/stable/

5. https://networkx.org/documentation/stable/index.html

6. Datasets used:

    1. https://networkrepository.com/fb-pages-food.php

    2. https://networkrepository.com/fb-pages-tvshow.php

    3. https://www.kaggle.com/datasets/wolfram77/graphs-arenas

THANK YOU