Proceedings of the International Conference on Smart Electronics and Communication (ICOSEC 2020)
IEEE Xplore Part Number: CFP20V90-ART; ISBN: 978-1-7281-5461-9

Recommendations in Social Network using Link Prediction Technique

Ramya BV
Department of ISE
B.M.S College of engineering
Affiliated to VTU
Bangalore, Karnataka
ramyabv34@gmail.com

Dr. N Sandeep Varma
Assistant Professor, Department Of ISE
B.M.S College of engineering
Affiliated to VTU
Bangalore, Karnataka
sandeepvarma.ise@bmsce.ac.in

R Indra
Assistant Professor, Department Of ISE
B.M.S College of engineering
Affiliated to VTU
Bangalore, Karnataka
indra.ise@bmsce.ac.in

Abstract— **Currently online social network with rapid development has become part of people's life. Based on user interest social network will change over time with different nodes and edges. Predicting new relation and missing relation between nodes in a social network can be identified using link prediction. Where new links and nodes can be identified with attribute information. Machine learning approaches are used with a set of features to increase the performance using supervised learning setup. Predicting the probability of connection between nodes for Twitch dataset, collected from SNAP and the main objective is to predict the probability of connection between nodes for personalized recommendations of game streamers using supervised machine learning by training model and performance of the model is evaluated using prediction performance metrics.**

Keywords— Social Network, Link Prediction, Machine learning, Performance metrics, Supervised learning, Twitch

I. INTRODUCTION

A social network represents interaction and relationship among the people in a group or community. A relationship can be of any social relationship such as friendship, family, colleagues, business partner and purchase. The network can be visualized as a graph with nodes and edges. Where individuals are represented by nodes and edges represents the association between nodes. Association can be formed with mutual interests in a community. Social network changes very quickly over time with establishing new links and also break of old links. Relationship between the nodes will also change over time. So predicting the missing relationship and future connecting links between node is a very important task in a social network.

Link Prediction in predicting links that either not yet exist at the given time or existing in each time t, but it is unknown at time t+1. Given a snapshot of a social network at some-time t, predicting links that will be formed newly in the network between the interval from time t to a given future time t+1. By using features of node attributes in the network, future links can be identified. Link prediction also can be used for recommendation systems in a social network, information retrieval and in many other fields. Probability of finding link formed between nodes is very important as social network structure varies over time. For finding links in the social network there are different methods of link prediction. These methods help in finding many feature set, which helps to increase accuracy.

Machine learning is a process of building a model to predict accurately based on data fed into the system. Different machine learning methods are used for building model. Where Supervised learning is used to train the model, which helps with better decisions and prediction.

Twitch is an online site which enables users to watch and broadcast the live streaming for gamers, allows gamers to chat and fans to connect with streamers. Used Twitch dataset from SNAP website where the dataset is stored, which is collected on May 2018. Here streamers are attempted to recommend based on binary classification with features recommend or do not recommend which is extracted from the dataset and trained the model using supervised machine learning algorithms to predict links between nodes and predicted links can be used for recommendation.

II. LITERATURE SURVEY

The survey paper by C.Mutlu et al [1] explained briefly about challenges in social network and approaching through machine learning methods with different Link prediction methods such as feature extraction methods and learning methods for predicting missing links and future possible connection between nodes with comparing each feature extracted. Dataset collection for link prediction and machine learning methods for dealing with the complex social network.

Mohammad J.Zaki et al[2] worked on a survey of link prediction in a social network, application on link prediction and ways to find features for different applications. The author has explained similarity-based and learning-based link prediction methods. In this, they have compared different link prediction methods for the large complex network by considering different models and trained model using binary classification, Baye's graphical models and similarity metrics.

Vineet Chaoji et al[3] worked with two different datasets and given a brief idea of using supervised machine learning for link prediction with identifying different features for solving easily. Model is evaluated using different performance

metrics such as accuracy, recall, precision, F1-score with cross-validation. Model is built using different machine learning algorithms and accuracy obtained above 80% for all models used. Among these algorithms, they found SVM with more accuracy of 90.56% and 83.18% compared to other models.

Jeyanthi et al [4] have proposed two-step solution for solving link prediction in a dynamic social network. Where the first step is constructing features with a combination of a domain and topological attributes of the graph, the second step is constructing features with the unconstrained edge. They have used similarity methods such as common neighbors, preferential attachment, Adamic-Adar, common keywords, and Degree mixing probability. Used logistic regression supervised learning algorithm.

III. PROPOSED METHODOLOGY

Aim of this project is to recommend a streamer with predicted links using link prediction methods with different features extracted from the dataset and using supervised machine learning algorithms. This helps users to connect with likeable streamers. To build a machine learning model, steps are followed as explained below:

- Data Collection
- Data Preprocessing
- Feature Engineering
- Test-Train-Split
- Building Classifier
- Evaluation

1. Data Collection

Twitch dataset collected from SNAP. Which is a user-user network of gamers Which consist of 7,126 nodes and 35,324 edges where nodes represent users and edges represents the friendship between users.

2. Data Pre-processing:

The dataset contains 35,324 edges but possible edges from the dataset are 50,772,750. To highly skew the dataset through missing edges 60,000 missing edges are randomly sampled by considering an edge as missing only if the shortest distance between the source node and destination node is more than 2. Where if the distance is less than 2 then there is a probability of connecting them in future. In case of distance more than 2 then the probability of connecting between nodes is less and those edges are considered as missing edges. Where edges already present in the network are represented with value 1 and missing edges are represented as 0(Figure 1). Therefore, classified edges with 1 as positive and edges with 0 as negative.

	Source	Destination	Class
0	6194	255	1
1	6194	980	1
2	6194	2992	1
3	6194	2507	1
4	6194	986	1
5	6194	4003	1

	Source	Destination	Class
0	40001	55650	0
1	46321	79731	0
2	6473	38286	0
3	71391	59506	0
4	75610	20001	0
5	75834	23159	0

Figure 1. Generating Class Labels

3. *Feature Engineering:*

Collected dataset has no features to work with. Therefore extracted different features from the dataset using different link prediction techniques and used top 10 features are selected from the Chi-Square test.

Following features are extracted from the dataset:
Jaccard Similarity Index: It calculates the similarity between the node pair. It is Jaccard similarity between nodes is mathematically represented as:

$$\text{Jaccard}(p,q) = \frac{(Np \cap Nq)}{(Np \cup Nq)} \quad (1)$$

Np and Nq represent neighbors of node p and q respectively.

Cosine Similarity Index: It measures similarity with the cosine of the angle between vectors which is pointing in the same direction. Cosine similarity between nodes is mathematically represented as:

$$\text{Cosine}(p,q) = \frac{(Np \cap Nq)}{\sqrt{(Np * Nq)}} \quad (2)$$

Frequency weighted common neighbour: Adamic Adar proposed this method with a smaller degree of common neighbours are weighted heavily. It is mathematically represented as following

$$\text{AA}(p,q) = \sum_{(z \in (Np \cap Nq))} 1/Log(Nz) \quad (3)$$

Page Rank: Algorithm used by Google search. It is used to measure the connectivity of nodes based on the in-degree of the node and it computes the ranking. Where page rank is used for both followers and followees.

$$\text{Pagerank}(i) = \sum_{(j,i) \in E} \frac{P(j)}{O(j)} \quad (4)$$

O(j) represents outer links of the node j.

Katz: It calculates all the paths between node pair with assigning a high score to the shortest path and low value to a longer path. It uses a factor of β to calculate values and mathematically represented as:

$$\textbf{Katz}(p,q) = \sum_{l=1}^{\infty} \boldsymbol{\beta}^{l} |\boldsymbol{paths}_{p,q}^{<l>}| \quad (5)$$

l is the length of all path between node pair.

Shortest Path: It is used to find the shortest path in the network. In our work direct edge between the source node and destination, node is removed because the shortest distance between them will be 1 and calculating the shortest distance for those nodes will not improve our model accuracy. Here, intuition is the shortest distance between nodes have a high probability of connecting in future so it can be used for recommendation.
Weakly Connected Components: It calculates all path between nodes without considering the direction.
Follow Back: If the source node is following the destination node and destination node is also following to source node then that nodes are added in this feature.

Inter Followers and Followee Count: It represents a number of common followers and followees between the source node and destination node.

Follower and Followee Counts: It is used to the calculated total number of followers and the total number of followees of source and destination node. The intuition of this feature is popular streamer has a greater number of followers and can be recommended for users.

Feature selection technique is the process of selecting the best features to reduce the overfitting, improve accuracy and reduce training time. Chi-Square test have been used for selecting top 10 features using KBest selection method.

Chi-Square: It measures the deviation between observed count and expected count for all features. The observed count will be close to the expected count when two features are independent. Value is calculated using a mathematical expression as:

$$x^2 = \sum \frac{(O-E)2}{E} \qquad (6)$$

0 is observed and E is expected values.

Top 10 features are selected using a Chi-square test are shown in figure 2.

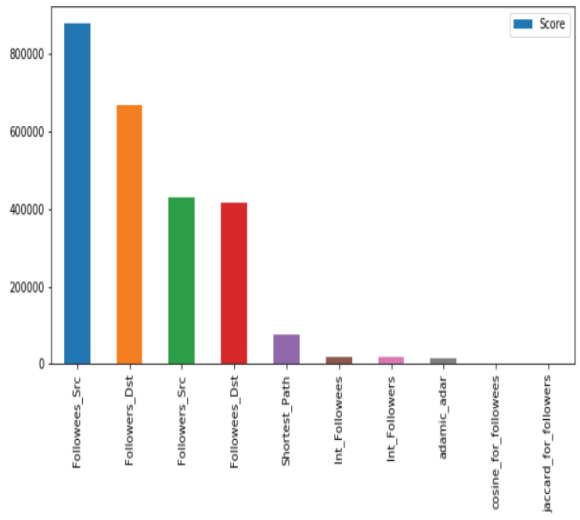

Figure 2. Top ten selected features.

4. *Test-Train Split*:

In machine learning dataset as to be split into train and test data. Where train data is used to train the model and model is tested with test data to measure the model performance. Splitting data is important because if the machine is trained and tested on the same data then measuring the accuracy of the model goes wrong. Therefore, any machine learning to be trained and test with different data. Therefore dataset is divided randomly both positive and negative data with 70% as train data and 30% as test data. Then using trained data, the model is trained and tested with test data.

5. *Building Classifiers:*

supervised machine learnings are used, where the model is trained with classification 1 and 0 representing connected edges and not connected edges respectively to build a model.

- ***Logistic Regression:*** It is a standard technique for social network analysis where it measures the relationship between the dependent variable which is labels(connected edges and not connected edges) and independent variables (features used). Probability of prediction is calculated using the below formula.

$$G(X)=ln\left[\frac{p(X)}{1-p(X)}\right] = \beta_0 + \beta_1 X \qquad (7)$$

- ***Bagging Classifier:*** It is one of the ensemble classifiers where prediction is calculated by aggregating individual prediction to reduce variance. Here dataset is divide into different samples and trained. Bagging classifier for decision tree is implemented with a majority vote and mathematical formula used as below

$$f(X)=sign\ (\sum_{i=1}^{T} sign(f_{i(x)})) \qquad (8)$$

- ***XGBoost:*** XGBoost refers to Extreme Gradient Boosting which is efficient for building a classification model and it is boosting algorithm which uses gradient boosting framework. XGBoost is highly scalable, quick to execute and efficiency is more compare to other machine learning models.

6. *Evaluation*:

Performance metric: After building the model, it is trained with train data and tested using test data. Now measuring the performance of the model is very important to know how model is performing in prediction. This can be done with using performance metrics which is explained below:

- ***Confusion Matix:*** Confusion matrix is used to find correctness and accuracy of the classification model. The output is represented in the form of a matrix and describes the performance of the model. Where each row represents an instance of the predicted class and each column represents an instance of an actual class. Four measures of confusion matrix are explained below.

Table 1. Confusion matrix

PV / AV	Predicted Positive	Predicted Negative
Actual Positive	TP	FP
Actual Negative	TN	FN

TP (True Positive) : The samples correctly predicted as future
links. Edge present between nodes is Positive (1), classified correctly as positive (1).

TN (True negative) : The samples correctly not predicted as future links. No edges between nodes is Negative (0), correctly classified Negative (0).

FP (False Positive) : The samples incorrectly predicted as future links Positive (1) edges are misclassified as negative (0) edge.

FN (False negative) : The samples incorrectly not predicted as future links. Negative (0) edges are misclassified as Positive (1) edge.

- **ROC Curve:** This performance metric is used for the binary classification model. It plots true positive rate(TPR) vs false positive rate(FPR) for various threshold values. Here Intuition of plotting is higher the ROC, the model predicts in better with predicting true positive and true negative. FPR and TRP is calculated as,

$$\text{TPR} = \frac{number\ \ of\ \ correctly\ predicted\ \ future\ \ links}{number\ \ of\ actual\ \ future\ links} = \frac{|TP|}{|TP|+|FN|} \quad (9)$$

$$\text{FPR} = \frac{number\ \ of\ \ incorrectly\ predicted\ \ future\ links}{number\ \ of\ actual\ \ negative\ links} = \frac{|FP|}{|FP|+|TN|} \quad (10)$$

- **Precision – Recall:** It is a plot between precision and recall. The high area under curve obtained from Precision-Recall curve represents high precision and high recall. where precision-recall is calculated as,

$$\textbf{\textit{Precision}} = \frac{TP}{TP+FP} \quad (11)$$

$$\textbf{\textit{Recall}} = \frac{TP}{TP+FN} \quad (12)$$

IV. RESULT AND ANALYSIS

The results of the proposed system are explained in this section based on the confusion matrix, ROC curve and Precision – recall curve. Using different supervised machine learning models data is trained and models are evaluated using performance metric. Confusion matrix values of Machine learning algorithms such as Logistic regression, XGboost and Bagging classifier values obtained from classifiers built are shown in Table 2.

Table 2 Confusion matrix result

Algorithm	TP	TN	FP	FN
LR	9711	17824	210	853
XGBoost	**10077**	**17869**	**487**	**165**
Bagging Classifier	10077	17821	487	213

From the Confusion matrix values obtained from different classifiers highest values of TP and TN have been obtained using XGBoost.

Using the performance values TP, TN, FP, and FN from the table 2 metrics precision, Recall, F1-score ad Accuracy are calculated. Below Table 3 shows the comparison of metrics obtained for machine learning algorithms.

Table 3. Precision, Recall, F1-score and Accuracy value comparison

Algorithm	Precision	Recall	F1-Score	Accuracy
LR	0.978	0.919	0.948	96.28
XGBoost	**0.983**	**0.953**	**0.968**	**97.72**
Bagging Classifier	0.977	0.951	0.963	97.55

From the Table 3 values obtained for Precision, Recall, F1-score and Accuracy scores obtained for XGBoost is more compare to logistic regression and Bagging classifier. XGBoost obtained an highest accuracy of **97.72%.**

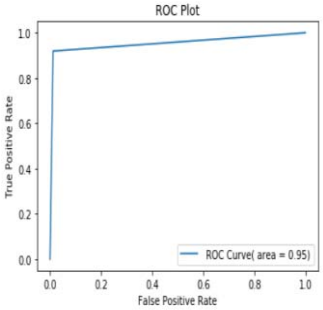

Figure 3. ROC plot for Logistic Regression

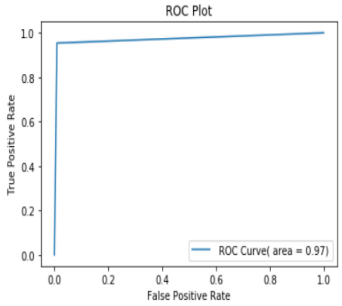

Figure 4. ROC plot for XGBoost

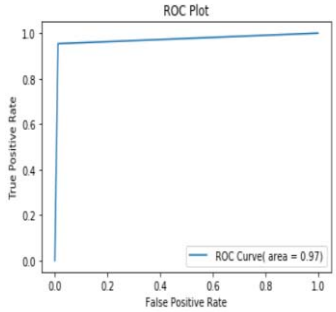

Figure 5. ROC plot for Bagging Classifier

ROC plot of Logistic Regression (*Figure 3*) obtained 95%

score. Highest ROC scores 97% is obtained for XGBoost (Figure 4*)* and Bagging classifier *(*Figure 5*)*.

Precision recall plot for machine learning algorithms is as shown below.

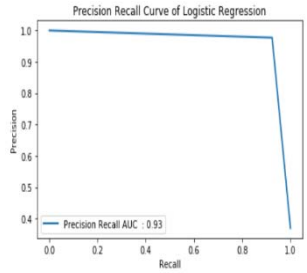

Figure 6. Precision – Recall plot for Logistic regression

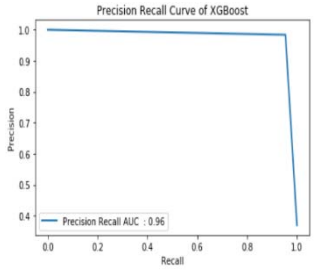

Figure 7. Precision – Recall plot for XGBoost

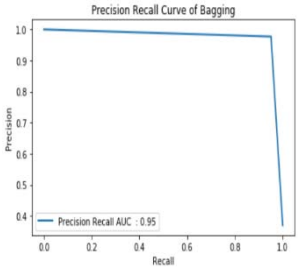

Figure 8. Precision – Recall plot for Bagging

Precision-Recall plot for logistic regression is obtained 93% (Figure 6), and 95% as been achieved by Bagging classifier *(*Figure 8*)*. XGBoost has obtained the highest Precision recall score as 96%(Figure 7).

From all the evaluation used obtained the highest score for XGBoost classifier when compare to logistic regression and Bagging classifier. Hence XGBoost is the best algorithm for recommending using different link prediction techniques.

V. CONCLUSION

In this work different machine learning models are used for detecting the probability of link formed in future between nodes which helps to recommend different streamers. Many unsupervised link prediction features are implemented and suitable features are chosen using chi-square test. Among Logistic Regression, XGboost and Bagging Classifier, XGBoost as obtained a good score in all the performance measures and obtained the highest accuracy of 97.72% compared to other implementation.

REFERENCES

[1] Ece C.Mutlu, Toktam A. Oghaz "Review on Graph Feature Learning and Feature Extraction Techniques for Link Prediction" , Jan 2019
[2] Mohanmmad AL Hasan, Mohammed J. Zaki "Link prediction in social networks"
[3] Vineet Chaoji, Mohammad AL Hasan, Saeed Salem, Mohammed Zaki, "Link prediction using supervised learning"
[4] Jeyanthi Narasimhan, and Lawrence Holder, "Feature Engineering for Supervised Link Prediction on Dynamic Social Networks".
[5] Abir DE, Soumen Chakrabarti, "Privacy preserving link prediction with latent Geometric netwrok models", 2019
[6] Layan Dong, yongli Li, Han Yin, Huang Le "The algorithm of link prediction on social network", Hindawi,2013
[7] Rediet Abebe and Vasileios Nakos, "Private Link Prediction in Social Networks", December 8,2014
[8] Yizhou Sun et al , "Co-Author Relationship Predictionin Heterogeneous Bibliographic Networks"
[9] David Libel-Nowell, Jon Kleninberg, " The link prediction problem for socail network"
[10] WANG Peng et al, " Link prediction in social networks: the state of the art",january 2015.
[11] https://www.zdnet.com/article/what-is-machine-learning-everything-you-need-to-know/
[12] https://www.zdnet.com/article/what-is-machine-learning-everything-you-need-to-know/
[13] Raf Guns, "Link Prediction" , Springer 2014.
[14] Jeyanthi Narasimhan, and Lawrence Holder, "Feature Engineering for Supervised Link Prediction on Dynamic Social Networks".
[15] Qi Ying Lim, Jason Wang, Ruifan Yang, "Social Network Analysis: Supervised Link Prediction", December 2017.
[16] Qi Yu1, Chao Long, Yanhua Lv, Hongfang Shao, Peifeng Hel, Zhiguang Duan4, "Predicting Co-Author Relationship in Medical Co-Authorship Networks", PLOS, July 2014.
[17] Manel Slokom, Raouia Ayach, "A New Social Recommender System Based on Link Prediction Across Heterogeneous Networks", Springer International Publishing AG 2018, 2017.
[18] Evan Darke, Zhouheng Zhuang, and Ziyue Wang, "Applying Link Prediction to Recommendation Systems forAmazonProducts".
[19] Mohamad AI Hasan, Mohammed J.Zaki, "Link Prediction in socail network".
[20] Liyan Dong, Yongli Li, " The algorithm of link prediction on social network", Hinawi, 17 sep 2013.