A Survey Of Link Prediction In Social Network Using Deep Learning Approach

Kalpana Prajapati, Harshal Shah, Rutvik Mehta

Abstract: In recent days social networks grow at every timestamp. Peoples are communicating with each other through social applications. So link prediction is the research trend in the current era. Through this link prediction, more people can connect with each other for communication-based on their similar features. In this area various machine learning techniques like supervised and unsupervised learning are used to improve the performance and similarities based features. Hence, Deep learning models are used to give better results as we take various analyses of research papers.

Index Terms: Deep Learning, Deep walk, Line ,Link prediction, Network embedding, Social network

1 INTRODUCTION
In the social network, peoples are connected with each other to do communication. The social network is a graph representation where two users are nodes and they associate with some form represent as an edge. There are undirected and direct links in the network graph. If there is a connection between sources to the destination, the link is directed and if there is no connection between sources to the destination, the link is undirected. In recent days, link prediction is the current trend to analyze the social network and predict future links which are the upcoming friends who are known or unknown in their mutual friends’ list. Many social network applications are used to predict the link such as Facebook, we-chat and LinkedIn. The social network is changing a period of time which is known as temporal network. In temporal large network is still a challenging task with largely recorded parameter responses to predict a link. The various link prediction methods are used to calculate the similarity of nodes, to predict a missing link, to detect a group of objects and the proximity of nodes. In early days, many efforts have been devoted to developing systems or tools to predict future links states based on historical data with state of the art methods which have some weakness like capacity and computational problems [1]. The supervised and unsupervised techniques are given a good performance to predict the link hence to improve the performance and to overcome these weaknesses various algorithms are designed to predict maximum relevance link prediction for the neighbor based networks in deep learning.

2 DIFFERENT METHODS OF LINK PREDICTION

2.1 Neighbor based Metrics
Neighbor based method is used to find similarity between two nodes.
1. Common Neighbors (CN): Given two nodes, x and y are connected with each other with a common neighbour [2]. It is calculated as: Common Neighbor (x, y) =|\Gamma(x) \cap \Gamma(y)|

2. Jaccard Coefficient (JC): It calculates as part of the common neighbours in the total number of neighbours [2]. It is defined as:
\[
\text{Jaccard Coefficient } (x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}
\]

3. Adamic-Adar coefficient (AA): Social network applies this technique in different situations. It is defined for less number of neighbors with heavy weights [2].
\[
\text{Adamic-Adar } (x, y) = \sum_{u \in N(x) \cap N(y)} \frac{1}{\log(|N(u)|)}
\]

2.2 Path-Based Metrics
The sequence of nodes is connected with the sequence of links in the network. Hence in this method nodes and neighbors information is required to calculate the similarity of nodes.
1. Shortest Distance (SD): If two nodes are more similar then their distance is short [2].
SD Score (x, y) = length (shorts(x, y))

2. Local Path (LP): There is an adjacency matrix at length 2 and 3 represent as A2 and A3. The adjacent factor could be fined as a which is the number between -1 \leq a \leq 1 [2].
LP = A2+aA3

3. Kartz (KA): This method is used when a path is long with less weight to find similarities. It is defined as a set of all paths between nodes and y and the growth of length decay factor is used between 0 \leq \beta \leq 1 [2].
Katz (x, y) = \sum_{i=1}^{\infty} \beta^i \times |\text{paths}_{k,i}(x,y)| = \beta A + \beta^2 A^2 + \beta^3 A^3 + ...

2.3 Random – Walk Based Metrics:
The source node to destination node random walk generates. It can be denoted by transitional probabilities between node and their neighbors. Based on a random walk, it can be calculated similarities between nodes through many link prediction metrics [2].

3 NETWORK EMBEDDING METHODS

3.1 Node2Vec:
Using deep learning, such applications are audio, video, network analysis and text analysis the Word2Vec and Node2Vec algorithms are there. For the document embedding and text embeddings, the Word2Vec algorithm is used. For the graphical structure and especially for the node embeddings the Node2Vec algorithm is used. By using the Node2Vec algorithm can be found maximum links in the many more applications like protein interactions, social network and co-authorship network. Firstly, the given datasets are plotted into...
a graph in a structured manner. Then the edge embedding algorithm is applied to predict the edge, similar edge and missing edge prediction from the two nodes. The random walks start from the source node and get the biased. To embed more information with a high standard biased random walk will be benefited [3]. They search the node from the graph-based algorithms. The random walk gives the best performance in time complexity. To learn the edge features there are many binary operators like Hadamard, Average, Weighted -L1, Weighted-L2. It gives a node assignment and often communication of the given node [4]. For the local structure, it is very good to develop a short random walk and improves the result. But in the global structure as the random walk generates a long path the performance may be degraded. For optimization, we can adopt a stochastic gradient descent method to improve better results.

3.2 Deep -Walk.
In the deep-walk, the graph is represented as a sparse matrix. Sequences of nodes are built up from a random walk. To learn the representation of nodes it uses a skip-gram model and improves a result. It gives an advantage over the implied representation of feature vector and sparsely labeled graph environment. For optimization, we can adopt gradient descent and back propagation [3].

3.3Line:
In the LINE algorithm, to optimize two independent functions for 1st order and 2nd order proximities the node’s feature vector is generated. It provides a final structure from the combinations of these functions [5].

4 DEEP LEARNING APPROACHES FOR LINK PREDICTION

4.1Deep Neural network
In DNN, the several layers are connected and produce the output of nodes where parameters are repeatedly tuned to minimize the error between actual output and final output. The sigmoid activation function is used to bind the score between 0 and 1 to find the similarities to predict a link. If the new time-consuming steps are added then it will give the worst result in the prediction process [6].

4.2 Deep Belief network
In DBN, the number of stacked RBM (Restricted Boltzmann Machine) is used for hidden activities training data. The RBM layers are iteratively trained and find the feature vector. The final feature vector learned by the top RBM’s hidden vector. Then after we can apply logistic regression to predict maximum likelihood links [7].

4.3 Deep Convolutional neural network
In a deep convolutional neural network, the connections are skipped to enter the layers without a selection of features. Beyond these layers, some information may be loosed. However, the deep convolutional neural network gives many short cut links to identify a mapping between nodes and gradient disappearance problem also solved [5]. Through the deep CNN the structural similarity of links we can get with better performance.

5 LITERATURE REVIEW
Chiu, Carter and Justin [6] proposed as link Prediction is the task to evaluate the existing link between the group of people or friends to find the future link between them. Many applications useful for link prediction like cyber security, biometrics, community structure, co-authorship structure. Link Prediction is a time-dependent problem, where network changes over time snapshot of the network graph. This paper presents the notion that a future edge between two nodes can be predicted with accuracy using proximity feature vectors. Here define the structure of the link to add and remove edges dynamically. For the weak estimators, stochastic learning principles to apply update probabilities to the new instances are observed with input vector to construct a training set to train a deep learning network. With the sigmoid function in the output layer to bound the score between 0 and 1 for single node. The datasets are Mathoverflow, EU-CORE, CollegeMsg used. Liu and Feng [7] analyze social networking websites that allow social members to interact with each other positively as well as negatively. Such interaction represented as a connection between them. The positive value means the agreement of interactions in the graph and a negative value means disagreement in the graph. Here paper presented to predict the viewpoint of one user toward form evidence-based on a relationship with different users of the surrounding social network. Based on deep learning approaches an unsupervised method to represent features for link prediction and link prediction method by deep neural network based on Restricted Boltzmann Machine (RBM). DBNS are learning a stack of RBM, there are three stack RBM used. Each of which is trained by using the hidden activities of previous RBM as its training data. Hence, the proposed method improves the link prediction using features with joint distributions. Janu and Verma [8], the proposed network embedding algorithm to recommend a friend for a large heterogeneous networks. It can be learned by edge embedding with multiple edge types into homogenous components. The method split the multigraph into a homogenous sub-graph with one edge type. The deep-walk and node2vec can be applied for the low dimensional space. Then, the neural network applies to train a heterogeneous edge embedding for link prediction. Over deep-walk, Node2vec neural net improves 84% accuracy. The friend recommendation dependent on this strategy is conveyed and presently running at Hike network application. Yue and Meng [9] proposed a model NeLSTM to predict link in the multi-link network and also proposed a time aging algorithm to divide a temporal network into a static network. LINE algorithm is used for dimensionality reduction and analyzed first order and second-order proximity connected. It gives high efficiency in large scale networks this model gives the solution of temporal link prediction. Chen and Jinvin [10], proposed supervised model for dynamic network link prediction methods to obtain better results. It used stacked LSTM into the encoder-decoder architecture to learn the representation of the graph of extracted features. It makes a long term prediction task for appearing and disappearing of links. It performed with the real-world datasets of dynamic networks. It implemented in the long term and short term networks. The model compared with cTRBM, GTRBM, DDNE, TNE methods. Thus the result can give better capture the pattern of network evolution in large scale networks also. It can be used in the local and global structure of nodes and beneficial for node classification. Keikha and Mohammad [11]
link prediction is to find the missing link on the social networks from the existing nodes and links. Link prediction is used for application such as to find interactions between proteins, recommendation systems, security domain, co-authorship network. Here, the paper presented a deep-link to extract features for structural and content information. For the structural information to consider path and community information and apply to classify WORD2VEC algorithm for finding the missing links in the given network to learn the Deeplink. For the whole content information extracted in one document using DOC2VEC algorithm. There may be a link between two users based on their interests and learned features of topics in the network. After learning these two pieces of information they concatenated and apply deep learning algorithms like NODE2VEC, DeepWalk, LINE, M-NMF. Compare AUC score with the different learning algorithms and neighbor based methods like common neighbor, Jaccard coefficient, preferential attachment, Resource allocation, Adamic Adar, Sorensen index. From the comparison, the DeepLink framework gives better performance for the limited resources. If the necessary structural and content information used then overall performance will be increased. So, the features are automatically extracted. For future work, it can be adopted different deep learning techniques to improve performance. For the criminal network, Lim and Marcus [12] proposed a model for missing and hidden links using supervised learning. In this approach deep reinforcement learning techniques adopted to predict hidden and missing links. It compares with the GBM model based on a binary classification task. The experiments indicate that the DRL method is more capable of better predictive performance. Li and al. [13] to predict edges according to historical link status that evolving each link’s pattern. It proposed GTRBM based framework to capture patterns of dynamic networks and integrates topological features and temporal attributes. The model simulates the transition patterns by using the adjacent matrix as input. It indicates that the prediction time reduced without accuracy reduction in real social networks. Thus it compares the framework against link prediction methods and shows that the GTRBM framework gives better accuracy. Link Prediction in a social network is to predict future links that will occur in the near future. Here, the paper [14] worked in the co-authorship network to analyze the features and metrics and assigned the score to all pairs of nodes then make the adjacency matrix. The co-authorship network node is the author and the link shows two authors have written at least a paper together. The similarity-based proximity score is measured. After calculating score using feature vector and dataset of co-authorship data for all pairs of nodes implement the artificial neural network. Form this paper also find low scored edges as weak links.

### TABLE 1

<table>
<thead>
<tr>
<th>Methodology</th>
<th>Dataset</th>
<th>AUC Score</th>
<th>Applicability</th>
<th>Scope capability inside</th>
<th>Limitations or challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNN [6]</td>
<td>Mathover flow</td>
<td>80%</td>
<td>Undirected, Dynamic network</td>
<td>8000 LINK</td>
<td>Challenge: High graph density &amp; strongly defined community structure</td>
</tr>
<tr>
<td>EU-CORE</td>
<td>Wikipedia admistration promotion</td>
<td>86% (logistic regression)</td>
<td>Undirected, Dynamic network</td>
<td>4000 LINK</td>
<td></td>
</tr>
<tr>
<td>DBN (3STACK RBM)</td>
<td>Wikipedia friendships</td>
<td>93%</td>
<td>Binary type</td>
<td>5000 Samples</td>
<td>Limitation: More RBM does not improve the result.</td>
</tr>
<tr>
<td></td>
<td>Infectious</td>
<td>92%</td>
<td>Multi-link</td>
<td>410 node</td>
<td>They are predicting similar features.</td>
</tr>
<tr>
<td></td>
<td>arXiv he-th</td>
<td>87%</td>
<td>Multi-link</td>
<td>22,908 node</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Highschool</td>
<td>95%</td>
<td>Directed positive weighted</td>
<td>70 nodes</td>
<td>It is based on local structure only improve the prediction.</td>
</tr>
<tr>
<td></td>
<td>Residence Hall</td>
<td>81%</td>
<td>Directed</td>
<td>217 nodes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Advogato</td>
<td>81%</td>
<td>Directed</td>
<td>6541 nodes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LDA [15]</td>
<td>92%</td>
<td>Directed positive weighted</td>
<td>70 nodes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Residence</td>
<td>82%</td>
<td>Directed</td>
<td>217 nodes</td>
<td></td>
</tr>
</tbody>
</table>

LISTR©2020

www.ijstr.org
6 CONCLUSION AND FUTURE SCOPE
From the above survey, the deep convolution neural network gives a better area under the curve (AUC) results than other techniques. The implementation of such a thing is based on the information available by the data collected from the various social networks. Deep learning-based algorithms are used to predict the future link and give better performance. Still some challenges in large scale networks to predict the future link and similarity between nodes. Hence we will improve some equations related to operators used in embedding algorithms. The various techniques are adopted and compare the result of the prediction.

REFERENCES