




RESEARCH ARTICLE

Supervised link prediction using structured-based feature extraction in social network

Anisha Kumari¹ | Ranjan Kumar Behera¹ | Kshira Sagar Sahoo²  |
Anand Nayyar³  | Ashish Kumar Luhach⁴  | Satya Prakash Sahoo¹

¹Department of Computer Science & Engineering, Veer Surendra Sai University of Technology, Burla, Odisha, 768018, India

²Department of Information Technology, VNR Vignana Jyothi Institute of Engineering & Technology, Hyderabad, 500090, India

³Graduate School, Duy Tan University, Da Nang 550000, Viet Nam, Faculty of Information Technology, Duy Tan University, Da Nang, Viet Nam

⁴Department of Electrical and Communication Engineering, The PNG University of Technology, Lae, Papua New Guinea

Correspondence

Ashish K. Luhach, Department of Electrical and Communication Engineering, The PNG University of Technology, Lae, Papua New Guinea

Email: ashishluhach@acm.org

Funding information

Veer Surendra Sai University of Technology (VSSUT), Burla, Odisha, India

Summary

Social network analysis (SNA) has attracted a lot of attention in several domains in the past decades. It can be of 2-folds: one is content-based, and another one is structured-based analysis. Link prediction is one of the emerging research problems, which comes under structured-based analysis that deals with predicting the missing link, which is likely to appear in the future. In this article, the supervised machine learning techniques have been implemented to predict the possibilities of establishing the links in future. The major contribution in this article lies in feature construction from the topological structure of the network. Several structured-based similarity measures have been considered for preparing the feature vector for each nonexisting links in the network. The performance of the proposed algorithm has been extensively validated by comparing with other link prediction algorithms using both real-world and synthetic data sets.

KEYWORDS

link prediction, similarity index, supervised learning, topological features

1 | INTRODUCTION

Social network analysis (SNA) has attracted a lot of attention to the researchers in the recent past years. SNA deals with an analysis of the information embedded in the social network. The information can be categorized into two types: first one is structured-based information, which represents the topological structure of the network, and the second one is content based, which represents the features associated with entities and their relationships. A number of research domains deal with linkage-based analysis, such as community detection,¹ network evolution,² link prediction, centrality analysis, and so on. Link prediction has several applications in various areas such as classification of harmful disease in health care, collaborative research prediction in coauthorship network, illegal connection mining in terrorist network, fraud detection and stock prediction in the financial sector, modeling recommender system for e-commerce industry,^{3,4} and so on.

The social network has become a common platform to share the information, ideas, text, multimedia data, and so on, which leads to the establishment of a number of relationships among the entities. Most of the links are established through social media platforms, especially when they are attracted toward a common interest. One real-time example is that friend suggestions that are popped-up when he/she belongs to a common friend in an online social network like Facebook. Similarly, some kind of product ads is popped-up on an e-commerce website when you are really having an interest in them. Link prediction deals with the identification of such relationships that are likely to be established in the future. However, predicting the relationships in the large-scale social network is found to be a challenging task due to its dynamic nature of the network structure, where links and nodes are added to the network over time. One of the major parameters that affect network dynamics is the type and strength of the relationship exist currently. Most of the social analyst is usually trying to make the link prediction based on the types of relationships in the present network and how they are going to affect others. It can be noted that the relationships could be among the individual entities, or it could

be between an individual and a community, or it could be between the communities also. The link prediction model may be based on two important factors. It could be based on the structural information like the way links are currently exist inside the network and the way structural equivalence is established among the entities. Sometimes, it also depends on the way the information is propagated in the network. The second kind of framework used to predict the links could be based on the feature associated with the individuals. The feature extraction step is one of the most crucial phases in this kind of model for predicting the nonexisting links.

A number of research works have been carried out to solve the link prediction problem by considering various similarity measures individually.⁵ It has been observed from the literature that any individual similarity measure is not enough to capture the link information in all the networks. In this article, we have considered several structured-based similarity measures to capture the strength of bonding in the network. We have proposed machine learning-based link prediction models in order to quantify the strength of relationships that may appear in the near future. The task of link prediction can be categorized as follows:⁶

- Missing link prediction
- Future link prediction

The missing link in networks is those which are not directly visible but are likely to exist through transitivity property. A link prediction model based on mining missing links highly depends on the amount of information that is hidden. In the case of missing link prediction model, someone can reconstruct a network of correlation taking into account observed fact and attempt to infer additional ties, which are not directly apparent, but likely to exist. However, in the case of the future link prediction model, one can consider the network evolution statistics to infer the future associations. Future link prediction is categorized in further two types; one is periodic link prediction wherein link is predicted in a timely manner, and the other one is nonperiodic link prediction, where the links are considered based on the particulars snapshot of the network.

The major contribution in this article is as follows:

- As the link prediction problem in a complex real-world network is computationally inefficient in nature, we have mapped the problem to the supervised machine learning domain, where the possibility of the appearance of future links is being predicted.
- The major issue in machine learning domain is the feature engineering. We have extracted the features for nonexisting links from the topological structure of the network by computing various local and global similarity measures. The actors who are having high similarity scores are more likely to be predicted.
- Instead of drawing a conclusion from any individual similarity measures, we have proposed machine learning models that combine all the structural features for link prediction.
- The different types of similarity measures that we have considered in this article are common neighbor (CN),⁷ Jaccard coefficient (JC),⁸ preferential attachment (PA),⁹ Adamic/Adar index (AI),¹⁰ Katz measure,³ SimRank,¹¹ and so on.

The subsequent sections of this article are presented as follows: In Section 2, the motivation toward machine learning-based link prediction models based on structured-based feature extraction has been discussed. Section 3 presents the literature survey on different approaches that have been adopted for link prediction. The methodology adopted in work is presented in Section 4. The proposed algorithm is presented in Section 5. In Section 6, the implementation of the proposed algorithm on both real-world and synthetic data sets has been discussed. A comparative study of different algorithms using evaluation metrics is presented in Section 7. The conclusion and future perspective of the work is presented in Section 8.

2 | MOTIVATION

The motivation toward the research work may be described as three folds:

- Link prediction is one of the widespread research discussions in the field of SNA. This problem could lead to a number of real-world applications like modeling recommender systems, fraud detection, stock prediction, and so on. If the missing links or the links to be appeared in the future are being predicted, one can understand the dynamics of network evolution. Based on the observed patterns of structural changes, one can definitely make better decision models for the business strategy, which can have high-performance value as well as less market risks.
- A number of research works in SNA have been focused on building an optimized model for predicting the links in real-world networks. Some of them have applied supervised machine learning algorithms by analyzing the content associated with nodes and edges. However, these are likely to have less accuracy as the content might have noisy and irrelevant data, which may require 90% of the time in data cleansing only. This motivates us to apply machine learning algorithms on features extracted from the structural information in the network. Structural information is less prone to noisiness.

- It has been observed that many authors have used similarity indices for predicting the future links.^{12,13} They have used only one of the structured-based similarity index in their algorithms. A wide number of similarity measures have been proposed based on a different aspect of structural information in the network. Therefore, an algorithm may have better performance in one similarity index, whereas it may not have remarkable performance in other similarity measures for the same network. So, the performance of the algorithms varies from network to network. This motivates us to use a series of similarity measures for predicting future or missing links. In this article, we have hybridized six different similarity measures to prepare the feature vector for missing links. Based on the feature vectors for all the nonexistence links, machine learning models have been designed, which are free from biases toward any particular structure-based information.

3 | RELATED WORK

The link prediction problem is found to be one of the emerging research directions in recent years. The first encouraging research work in this area was carried out by Liben-Nowell and Kleinberg, where they observed that the topological information is highly effective for predicting future relationships in the network.³ Their work emphasizes on modeling the network evolution by leveraging the topological information in the network. They have proposed various similarity measures based on node neighborhood and graph distances for link prediction. A trade-off between accuracy and computation cost is associated with each of these similarity measures, which are based on proximity evaluation methods. Some of the elegant works in the area of link prediction are presented in Table 1. Newman has proposed a method in which links are predicted by covering all the paths of restricted length based on small-world hypothesis.²¹ The idea behind the hypothesis is that one person can be reached to another person through several paths of different lengths and two persons in any large network can be connected through a small number of acquaintances.^{22,23} A small world network follows power-law distribution in terms of both degree and edge. It is found to be much popular in social network analysis as it always resembles the real-world networks, which are scale free in nature. Power-law degree distribution can be defined as defined in Equation (1).²⁴

$$N(d) \propto d^{-\gamma}, \quad (1)$$

where $N(d)$ is the number of nodes having degree d and γ is the scale-free exponent for power law distribution. Some of the research works have also argued that the link prediction problem belongs to the class of NP-hard problem.²⁵ It may be observed from literature that positive links play a crucial role in network evolution in signed social network.²⁶⁻²⁸ Individuals maintain positive affections as well as negative affections relative to one another, so social network functions as a platform to exhibit this type of relationship, whether healthy or unhealthy, friendly or hostile, like or unlike, belief or disbelief. These kinds of relationships give rise to the appearance of signed social networks (SSNs) in which positive sign means like a friend, trust, and negative sign presents dislike, rival, and distrust.²⁹ That is, social networks contain information regarding the positive links or negative links that highly influence the establishment of the future structure of the network. However, most of the social webs considered as possessing only friendly (positive) relationships, whereas neglecting the hostile (negative) ones entirely.

Yuan et al.³⁰ have proposed a graph kernel-based method to infer the future links in signed social network. They believe that signed structural information in a network plays a major role in social evolution. In their work, they have generated many subgraphs for each user, which

TABLE 1 Some of the pioneer works in the area of link prediction

Author	Year	Technique adopted	Performance metrics	Structured/content based
Liben-Nowell and Kleinberg ³	2004	Proximity of nodes	AUC	Structured
Al Hassan et al. ⁴	2006	Supervised learning	Accuracy	Content
Lu and Zhou ⁶	2011	Similarity or node proximity	Accuracy	Structured
Papadimitrio et al. ¹⁴	2012	Friend link algorithm	AUC , precision, recall	Structured
Dong et al. ¹⁵	2013	CNFG algorithm and KatGF algorithm	ROC	Content
Gao et al. ¹⁶	2014	Neighborhood similarity	AUC	Content
Sadeque et al. ¹⁷	2015	Machine learning	Accuracy	Content
Yao et al. ¹⁸	2016	Network Evaluation matrix	ROC , AUC	Structured
Haghani and Keyvanpour ¹⁹	2017	Scoring based method & Deep learning	ROC , accuracy, precision	Content
Zhou et al. ²⁰	2018	Similarity based	ROC	Structured

depend on the strength of relationships and then compute the similarity using Bhattacharjya kernel. Finally, they have used SVM classifier to predict future associations. Yaghi et al.³¹ have presented a model based on an evolutionary neural network for predicting future links. They have focused on addressing class imbalanced distribution by using the random and undersampling method. In their work, they have adopted three optimizers, such as swarm optimization, moth search, and genetic algorithm for training feed-forward neural network. Alzubi³² have proposed several models, which focused on optimizing various ensemble classifier for better accuracy. Their optimization models are mainly focused on game-theoretic approach,³² consensus based,³³ and diversity based.³⁴ These ensemble approaches can be utilized for improving the performance of link prediction models.

Machine learning has become a center of attraction in various tasks of classification and prediction. In the area of links prediction, most of the works based on the similarity measures for the node pairs, which are found to be unsupervised methods. The unsupervised methods can be extended to supervised binary classification in the area of link prediction, where the links are being labeled at the time of data set preparation. Liben-Nowell and Klenberg³ have presented a path-based link prediction algorithm where the probability of the appearance of links depends on the length and number of the shortest path between pair of nodes. They have also presented an extensive comparison between node-based and path-based similarity measures. Since then, this article has become a benchmark for researchers who are working in the field of link prediction.

Al Hassan et al. has considered a supervised machine learning approach for the prediction of nonexistent links in their article.¹⁶ They have developed several machine learning models to capture the topological information associated with links and nodes in the network. They have extracted the proximity feature based on the content associated with the nodes. In their article, SVM is found to outperform other machine learning models. They build up the model based on proximity features under the supervised learning methods. Their work includes different supervised learning method, out of which SVM has better performance. Later Lu and Zhou⁶ in their survey summarized the development in the link prediction algorithms along with the contributions of physicists. Their work emphasizes the random walk and maximum likelihood methods. In their work, they have qualified the strength of bonding through a connection-weight score for each node pair. They have modeled their classifier based on the identified weight-score.

Papadimitriou et al.¹⁴ proposed a model for identifying the friend suggestions, which is based on the link prediction. In their work, the friend link algorithm is adopted, where small-world hypothesis has been considered. They have also adopted the Map-Reduce programming model for implementation in Hadoop platform in order to tackle huge size network. Liyan Dong et al. has used the DBLP experimental data sets, which is based on the multiple attributes extracted from the local and global structure in the network.⁶ Both the local and global similarity indices are applied based on node guidance capability. The six degrees of separation effect in the social network has been used for link prediction in work carried out by Fei Gal et al.⁸

A pioneer work based on a machine learning model applied on weighted real-world networks has been carried out by Sadeque et al.¹⁷ They have used group-based social network, that is, online health forum, where individuals are sharing their personal experiences and advice on a single platform. Their work is emphasizing on predicting whether or not a user will continue as a member of this group. Their model with 1-month observation predicts continuous membership with over 83% accuracy, which is later increased when 12-month data have considered for training. Later on, Lin Yao et al. proposed a model by considering network evaluation over a time period.³⁵ By considering the dynamic feature of social networks, they have considered three matrices, such as the time varied weight, the change of degree of common neighbor, and intimacy between common neighbors. They reexamined common neighbor by considering them within two hops, as the important task of link mining. Haghani and Keyvanpour have extended the link prediction problem from the similarity index based methodologies to the deep learning approach.¹⁹ They have considered dynamic behavior of social network, which depends mainly on two important features such as node information and linkage information about relationship between two nodes. Kai Zhou et al. focuses on similarity score-based link prediction by using two broad classes of approaches, including local- and global structure-based methodologies.²⁰

Some of the researchers have focused on link prediction in social networks using the features of users obtain from social media data. Social Internet of things plays a key role in identifying the link prediction in a large-scale complex network. As the data size is huge and complex, traditional tools fail to process social media data. Ahmed et al. has proposed a big data model for removal of noise and feature selection from the social media data, which can further utilized for predicting future associations.^{36,37} Paul et al.³⁸ have presented a model known as smart buddy, which analyzes human behavior from the social data obtained from the network of devices in IoT. This approach is definitely helpful in modeling the evolution of social networks by analyzing the link establishment.

Integration of smart systems and the IoT devices has a great influence on lifestyle of modern society.³⁹ However, the major issue in processing a huge amount of big heterogeneous data captured in smart devices is its scalability and computational cost. Rathore et al. have proposed a parallel distributed system for filtering, processing, and classifying a huge amount of data.⁴⁰ Din et al. have presented a framework for mobility management for a group of sensor nodes.⁴¹ The mobility management in sensor networks could be an application for link prediction in wireless mesh networks. Behera et al. have proposed a distributed big data model to capture the topological information for link prediction.⁴² In their work, they have adopted Hadoop framework to process large-scale complex networks. Paul et al. have presented a survey on the role of IoT in mining future association in social networks.⁴³

4 | METHODOLOGIES AND BACKGROUND DETAILS

4.1 | Structured-based similarity index

In this section, we have presented some of the useful methods for the prediction of nonexistence links, which are likely to appear in the future. The prediction of links is based on a similarity score, which is measured by considering either local or global information in the network. This concept allows us to estimate the degree of similarity between the pair of nonexisting nodes in the graph. In some traditional approaches, the link predicted between pair of nodes solely depends on their position in the network, and in some cases, it depends on the network structure. The former approach is termed as a local approach as it considers the local positioning of nodes in the graph, while the latter approach is termed as a global approach as it considers the position of a node in a network providing a wider view of similarity.

4.2 | Node-based similarity measure (local approach)

Let G denotes a graph representing a social network at time t_1 . $N(x)$ and $N(y)$ are neighbors set of x and y in G , respectively. Many link prediction heuristics assume that the probability of existence of future link between x and y at t_2 ($t_2 > t_1$) is more, if they share more number of common neighbor. To be more specific, x and y are more likely to be connected in future, if they are having more number of common neighbors. Based on this assumption, a number of metrics have been proposed in order to quantify the hidden links between the node pairs. Some of the important metrics are identified below:

4.2.1 | Common neighbors

The basic idea behind common neighbors is that the node pairs are more likely to establish link in future if they are sharing more number of common neighbors. This measure can be derived from the transitive properties of the social network. It can be defined as follows.⁷

$$LP_{CN}(x, y) = |N(x) \cap N(y)|, \quad (2)$$

where $LP_{CN}(x, y)$ is the predicted score of hidden link between node x and node y using common neighbors. In case of weighted graph, weight of links between nodes and their common neighbors may increase the predicted score. It can be mathematically represented in Equation (3)

$$LP_{CN}(x, y) = \sum_{z \in (N(x) \cap N(y))} w(x, z) + w(z, y), \quad (3)$$

where $w(x, z)$ is the weight of link between node x and node z . Node z is the common neighbor of node x and y .

4.2.2 | Jaccard coefficient

Jaccard coefficient gives the normalized score to the common neighbor's predicted score. It gives the probability score for hidden links between two nodes. Therefore, score is normalized between 0 and 1 and the expression can be represented as given in Equation (4).⁸

$$LP_{JC}(x, y) = \frac{|N(x) \cap N(y)|}{|N(x) \cup N(y)|}, \quad (4)$$

where $LP_{JC}(x, y)$ is the predicted score of hidden link between node x and node y , using Jaccard coefficient. Here denominator indicate the maximum possible common neighbors between node x and y .

4.2.3 | Preferential attachment

It has been observed that the chances of establishing links between nodes having higher degree is more compared with pair of nodes having smaller degree. The intuition behind this measure is that nodes which are highly connected in the network are more likely to establish new relationships with other nodes. The predicted score for hidden links can be measured by multiply the degree of both the nodes. It can be mathematically represented as follows:⁹

$$LP_{PA}(x, y) = N(x) * N(y), \quad (5)$$

where $LP_{PA}(x, y)$ is the predicted score of hidden link between node x and node y using preferential attachments.

4.2.4 | Adamic Adar

In Adamic Adar, the node pairs are assigned with a high score if their common neighbors are not being shared with other nodes.⁴⁴ This measure is based on the fact that a pair of nodes have a higher tendency to establish future links if they are connected to a common neighbor, which itself has no further relationships with other nodes. More is the value of such common neighbors, better is the chances for establishing links between said nodes.¹⁰ This can be mathematically expressed as given in Equation (6).

$$LP_{AA}(x, y) = \sum_{z \in (N(x) \cap N(y))} \frac{1}{\log(N(z))}, \quad (6)$$

where $LP_{AA}(x, y)$ is the predicted score of hidden link between node x and node y using Adamic Adar and z is the adjacent node of both x and y .

4.3 | Path-based similarity measure (global approach)

A number of approaches have been made to predict future connections between two nodes based on the present configuration of the network. The assumption behind the prediction is that if two nodes are going to be connected in future then there has to be a path between these nodes. The more and shorter these paths are, the better the chances for them to be get connected. Based on this idea, there are many heuristics like Katz measure, PageRank, and SimRank.¹²

4.3.1 | Kartz measure

This is the most useful measure to quantify hidden links between two nodes, which is based on the number of paths between the nodes. Here all paths having length less than diameter have been considered. This measure is calculated as the summation of a total number of paths of all length existing between a pair of node. It can be defined as³

$$LP_{KZ}(x, y) = \sum_{i=1}^{\infty} \beta^i |\text{path}_{x,y}^i| \quad (7)$$

where $LP_{KZ}(x, y)$ is the link predicted score using Karz measure, β is the small constant value, and $\text{path}_{x,y}^i$ is the set of all path lengths from x to y .

4.3.2 | SimRank

This measure considers the similarity score between the neighboring nodes. In this measure, the probability of establishing links for a node pair is high, if they are having more number of similar neighbors. It can be defined in Equation (8).¹¹

$$LP_{SR}(x, y) = \gamma \frac{\sum_{a \in N(x)} \sum_{b \in N(y)} LP_{SR}(a, b)}{|N(x)| |N(y)|}, \quad (8)$$

where $LP_{SR}(x, y)$ is the similarity score between x and y using SimRank. γ is the real number lies between 0 to 1.

4.4 | Machine learning algorithms adopted

A number of machine learning models can be useful in predicting the links in social networks. We have implemented some of the machine learning models in our proposed model, which are listed below:

- **Support vector machine (SVM):** Support vector machine is a classification method that is nonprobabilistic in nature. Sometimes it is also known as support vector network. SVM is the combination of supervised learning methods that are applied for both classification and regression. In

other words, for a given set of training data, the algorithm set up a framework that classifies a new example comes down into one of two categories based on the features of that data.

- **Decision tree (DT):** Decision tree is one type of supervised machine learning algorithm that uses a tree-like structure as a dividing model to go from observations about an item from root to the conclusion that is leaves. Here discrete values are assigned to the target variables.
- **Artificial neural network (ANN):** It is a computation system based on the analogy of biological neural network, where the building block is nerve cell (called neurons) used to process the information. Based on the requirement, the ANN model is partitioned into different layers, and each layer contains a set of nodes that are connected with nodes of the previous layer. The connections have different weights upon them. The feedforward network and feedback network are different variations of neural networks. But, in the feedback network, the loop is present so that according to error, the system parameters are changes continuously until it reaches a state of equilibrium.
- **Random forest (RF):** The number of decision trees is collected to form a random forest that can be applied to both classifications as well as a regression problem. Here the more the number of trees in forest implies more accurate results. For a large data set, the number of variables is more, and it is not easy to cluster the data, so this RF method can give better accuracy for data that belong in a particular group. For different trees, the different variables decided randomly. The tree which gives the best classification is predicted by using the testing data set.
- **K-nearest neighbor (KNN):** It is one of the supervised machine learning algorithms, which is quite simple and easy to implement. KNN can be applied to solve both classification and regression problems. The basic assumption of K-NN algorithm is that similar things are closer to each other. The distance between the data points is calculated mathematically euclidean measure or any other measures. Here, initialization of k depends on our chosen number of neighbors.

5 | PROPOSED ALGORITHM

Machine learning algorithms are found to be suitable candidates to solve link prediction problem in social network analysis. In this article, we have proposed a structured based machine learning models for link prediction, where both missing links or the future links that are likely to appear are being predicted. We have mapped the link prediction problem as a supervised binary classification problem where each missing linked is labeled as either predicted or not predicted in future. The execution flow of the proposed model is presented in Figure 1.

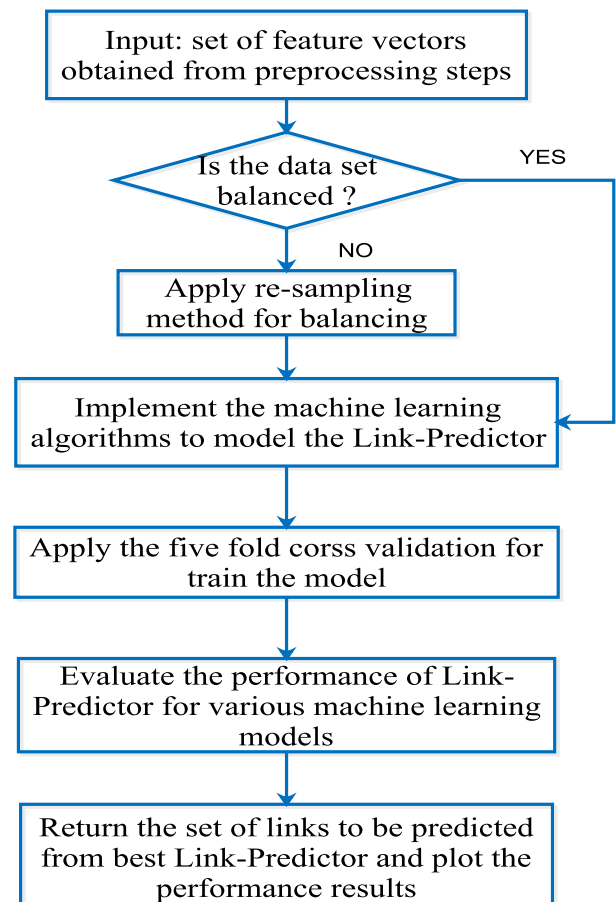


FIGURE 1 Execution flow of proposed approach for link prediction

The working flow of the proposed framework can be presented with the following steps.

- The four real-world and three synthetic data sets, in the form of edge-list format, are considered for the experiment. In order to process it in machine learning models, we have first transformed the network into a list of feature vectors, where each vector consists of a set of features for a nonexistent edge.
- The feature vectors have been constructed using various similarity measures as available in the literature. The preprocessing of the data set has been briefly discussed in Section 5.2.
- While distributing the classes within a data set, the major problem arises in machine learning is the huge margin in number of data between two classes, that is, data set is said to be unbalanced if the total number of data having positive class is far less than a total number of data having negative class. There are various techniques to solve this class imbalance problem like resampling (over sampling and under sampling), Ensemble methods, clustering the abundant class. We have considered the resampling method to balance the data set if it is unbalanced in nature.
- We have mapped the link prediction problem as the supervised learning model, where several machine learning algorithms have been applied to predict the future links.
- The link prediction model has been trained using 5-fold crossvalidation, where the data are partitioned into 5-folds, and each fold is used for validating the model.
- The performance of the proposed model is then evaluated using metrics like AUC, accuracy, F-measure, precision, recall, and so on.

6 | IMPLEMENTATION

6.1 | Data set considered

We have considered the networks from various domains having different properties. In all networks, we have not considered the multiedge and self-loop. We have experimented with the proposed model on four real-world networks and three synthetic networks. The Dolphin, Facebook NIPS, and Hamsterster friendship are collected through the Koblenze Network Collection (KONECT). The email-Eu-core have been obtained from Stanford Large Network Dataset Collection (snap-stanford). The detailed statistics of the data sets are presented in Table 2.

In this article, we have considered the topological structure of a static network for computing the similarity value for predicting future associations. As the structural information is less prone to noise, we have considered only the unweighted graph. We have not considered the strength of the relationship, which may be obtained from the weighted graph or the signed network. For the sake of simplicity and better performance, we have considered the unweighted graph only. In the future, this work can be extended to the weighted network, which represents the strength of social bonding among the users.

- *Dolphin*: This network is an undirected and unweighted social network of bottlenose dolphins. The date of the origin of dolphins is in between 1994 and 2001. In this data set, the node represents bottlenose dolphins of bottlenose dolphin community, and an edge represents frequent associations between them. This network was made available by Lusseau et al. in 2003.⁴⁵
- *Hamsterster Friendship*: This data set is the network of friendships among the users of the websites, hamsterster.com. In this network, the node represents users, and an edge represents the friendship.⁴⁶

TABLE 2 Data set considered for the experiment

Sl. No.	Data sets	Nodes	Edges	Clustering coeff.	Average degree	Directed/undirected	Balanced/imbalanced (after feature vector construction)
1	Dolphin	62	159	0.309	5.129	Undirected	Balanced
2	email-Eu-core	1005	25 571	0.399	4.261	Directed	Balanced
3	Facebook NIPS	2888	2981	0.0359	2.0644	Undirected	Imbalanced
4	Hamsterster Friendship	1858	12 534	0.0904	13.494	Undirected	Imbalanced
5	Synthetic network 1	5000	15 000	0.55	10	Undirected	Balanced
6	Synthetic network 2	10 000	30 000	0.346	15	Directed	Imbalanced
7	Synthetic network 3	15 000	50 000	0.336	12	Undirected	Imbalanced

TABLE 3 Parameters used for generating synthetic networks in LFR benchmark

Data set	N	E	d_{Avg}	d_{max}	μ	γ	β	s_{min}	s_{min}
Synthetic network 1	5000	15 000	10	30	0.3	0.55	-2	30	50
Synthetic network 2	10 000	30 000	15	35	0.2	0.346	-1.5	35	65
Synthetic network 3	15 000	50 000	12	50	0.2	0.336	-5	50	100

Abbreviations: β , exponent of power law degree distribution; γ , clustering coefficient; μ , mixing parameter; d_{Avg} , average degree; d_{max} , maximum degree; E: Number of edges; N, number of nodes; s_{max} , maximum community size; s_{min} , minimum community size.

- *Facebook NIPS*: This network is an undirected and unweighted social network that represents user to user friendship. In this network, the user is represented by nodes, and friendship is represented by the associated link between two nodes. This network was made available through Julian McAuley and Jure Leskovic in 2012.⁴⁷
- *Email-Eu-core*: This is an e-mail-communication network generated by the members of a large European institution. In this network, node represents the users and edge is established between two nodes, if at least a mail is generated by one of the two nodes. This communication through email only happens among institutions members, and the data set does not contain an incoming message from or outgoing message to the rest of the world. This network was made available through Hao Yin et al. in 2017 and through Leskovec et al. in 2007.⁴⁸

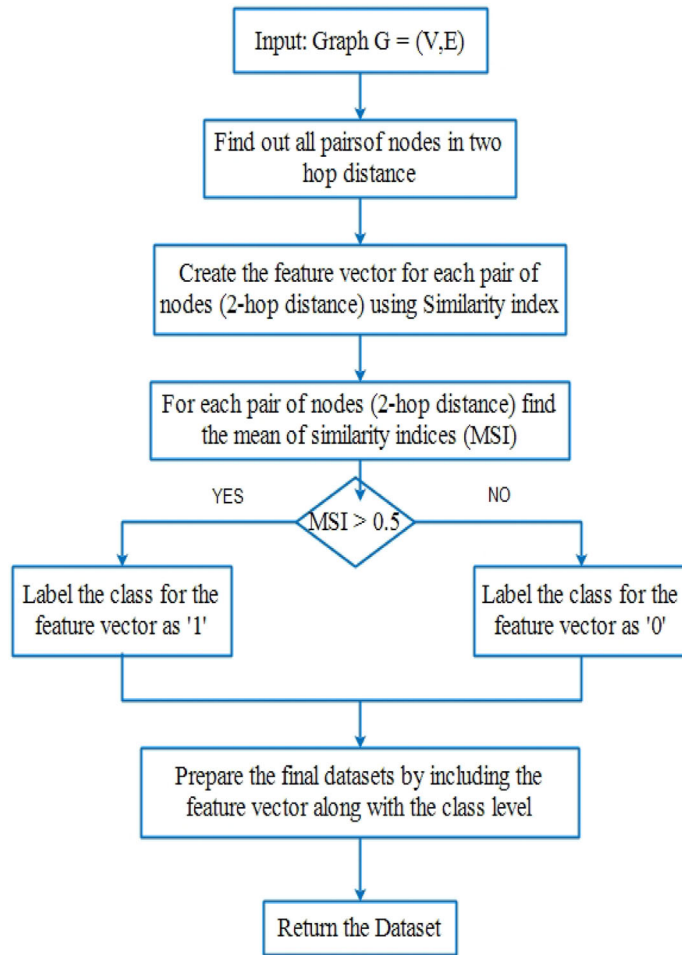
Apart from the real-world networks, three synthetic data sets have also been considered for the experiment, which have been generated using LFR benchmark. We have set the parameters of network in such a way that the network generated seems to resemble with real-world networks. The parameters that have been considered to generate the artificial networks are presented in Table 3. We have fixed up various parameters to generate artificial networks such as the number of nodes, number of edges, average degree, maximum degree, mixing parameter, clustering coefficient, exponent of the degree distribution, minimum community size, and maximum community size. The mixing parameter controls the edge density among the communities.

6.2 | Feature vector preparation

We have constructed the data set using various topological similarity metrics, as discussed in Section 4.1. Each feature of the data set is one of the similarity measures. As the social network is observed to be large in terms of complexity and size, we have considered all the links that are at two distance from each other. We have predicted the link score only for the links which are at two-hop distance. The assumption here is that the nodes which are away from each other at two distances are more likely to be connected in the future. The flow diagram for data set preparation is presented in Figure 2. The steps followed in the data set constructed are described as follows.

1. The network we have considered is first is transformed into edge-list format in preprocessing stage.
2. For a given network, huge number of pair of nodes may possible between which link does not exist at current time. To process all the nonexistent pairs may be infeasible. In order to reduce the computational complexity, we have considered only such pair of nodes, which are away in two-hop distance. We have extracted the features for all those pairs of nodes, which are at two-hop distance only.
3. After getting the node pairs having two-hop distance, the feature vector for each pair of nodes (two-hop distance) is constructed. The feature vector is constructed by computing various similarity measures as discussed in Section 4.1. In our proposed work, we have considered common neighbor (CN), Jaccard coefficient (JC), preferential attachment (PA), Adamic Adar (AA), Katz measure, and SimRank as similarity measures in feature extraction process.
4. After feature construction in step 3, we have labeled all the node pair to map the problem into binary classification. In order to label the feature vector corresponding to each edge, the mean of similarity indices (MSI) has been calculated by taking average of all the structured-based similarity that we have considered in previous step.
5. In this step, each of the feature vector has been labeled. The pair of node for which MSI value is greater than 0.5 is labeled as one and the pair of nodes having MSI value less than 0.5 is labeled as zero.

The feature vector is constructed for nonexistence links using six different similarity measures, which are computed from the topological structure of the network. Out of six features, four of them are local similarity measures and two of them are the global similarity measure. The local similarity measure takes account of the structural information around the nodes whereas the global similarity measure considers the path-based information, which includes both the number and the length of the shortest paths between a pair of nodes. The different similarity measures that we have adopted in the article are common neighbors, Jaccard coefficient, preferential attachment, Adamic Adar, Katz measure, and SimRank. The first four, that is, common neighbor, Jaccard coefficient, preferential attachment, Adamic Adar are the local similarity measure. We have considered Katz measure and SimRank as global similarity measures.

FIGURE 2 Flow diagram for data set preparation

Some of the data sets obtained from feature extraction process are found to be unbalanced in nature. We have adopted resampling approach for balanced the data set. In resampling approach instead of taking all the instances for train the model, different ratio of rare to abundant classes have been considered for balancing the data set. After the data set is sampled for processing, 5-fold crossvalidation has been applied.

7 | EXPERIMENTAL RESULTS AND DISCUSSION

7.1 | Experimental setup

All the experiment have been performed on a i7 processor having 3.4GHz clock rate. We have used igraph package of R language to measure the structural similarity among the nodes in the network. All the codes have been executed in R language, However, Matlab 2018b has been considered for plotting the graph for visualization purpose.

The proposed machine learning models have been compared with standard link prediction algorithms. The performance of the models have been evaluated using accuracy, F-score, precision, and recall. We have also adopted the AUC value to measure the performance.

7.2 | Performance evaluation metric

As the proposed approach for link prediction has modeled into a binary classification problem, the performance of the link prediction model has been evaluated with the help of confusion matrix, which is generated after the classification process. Some of the terms related to confusion matrix are discussed as below:

- TP: number of links that are labeled as predicted and also actually predicted by the link predictor.
- FP : number of links that labeled as not to be predicted but are being predicted by the link predictor.

- TN : number of links that are labeled as not to be predicted and are actually not predicted by the link predictor.
- FN : number of links that are labeled as to be predicted and are not predicted by the link predictor.

The evaluation metric that is used in our proposed work is discussed as below:

- Accuracy: The accuracy of the model is defined as the ratio of number of links that are actually predicted correctly to the total number of nonexistence links (only two-hop distances).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

- Precision: Precision is the ratio of TP to the summation of TP and FP.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (10)$$

Here the denominator is the total number of links that are predicted by the model.

- Recall: Recall is the ratio of TP to the summation of TP and FN.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (11)$$

Here the denominator represents the number of links that are labeled as to be predicted in the data set.

- F-score: F-score is the harmonic mean between precision and recall and its score ranges from 0 to 1. The mathematical representation for F-score is defined as below:

$$\text{F-score} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \quad (12)$$

- ROC curve: ROC is receiver operating characteristics and the curve is plotted against true positive rate (TPR) and false positive rate (FPR) for various threshold values. Here TPR is same as the recall as discussed in Equation (10). FPR is defined as below:

$$\text{FPR} = \frac{FP}{FP + TN} \quad (13)$$

7.3 | Results analysis of real-world network data sets

The comparative analysis in term of different evaluation measures has been presented in a series of graph from Figure 3A to Figure 4D. Figure 3A presents the comparative analysis in terms of accuracy on four real-world network data sets. The performance of various approaches adopted in this article for link prediction includes both machine learning and individual similarity measures. It can be observed that machine learning algorithms provide better accuracy compared with individual structured-based link prediction algorithms. SVM is found to have better performance compared with other algorithms in Dolphin, Facebook, Hamsterster friendship data sets. However, random forest gives better accuracy in the email communication network. Adamic Adar (AA) and common neighbor (CN) are found to have similar accuracy in Dolphin, Facebook, and email communication networks. It can be observed that although decision tree has lesser accuracy compared with other machine learning models for link prediction, it is better than all the individual structured-based link prediction.

Figure 3B presents the comparative analysis of link prediction algorithms in terms of precision in real-world networks. SVM has better precision value in real-world data sets, especially in case of large size data sets like Facebook, Hamsterster, and email communication networks. It can be observed from Figure 3B that the Adamic Adar algorithm has less precision value in Dolphin data set. JC and Karz algorithm has the lowest precision value in Facebook and Hamsterster data set, respectively. However, in the email communication network, the Common Neighbor seems to have the lowest performance in terms of precision. As the performance highly depends on the structural topology of the networks, the pattern of performance of these algorithms is quite difficult to predict. However, from the observation, it can be concluded that machine learning-based algorithms are having better performance in terms of precision and accuracy.

Figure 3C presents the comparative analysis of link prediction models in terms of recall. Machine learning algorithms are found to have better performance with less margin for small-sized data sets compared with larger data sets like Facebook, Hamsterster, and email communication networks. In the case of email communication network data set, the decision tree is found to have the best performance in terms of recall. However, SVM has the best performance on Facebook, Dolphin, and Hamsterster network data sets.

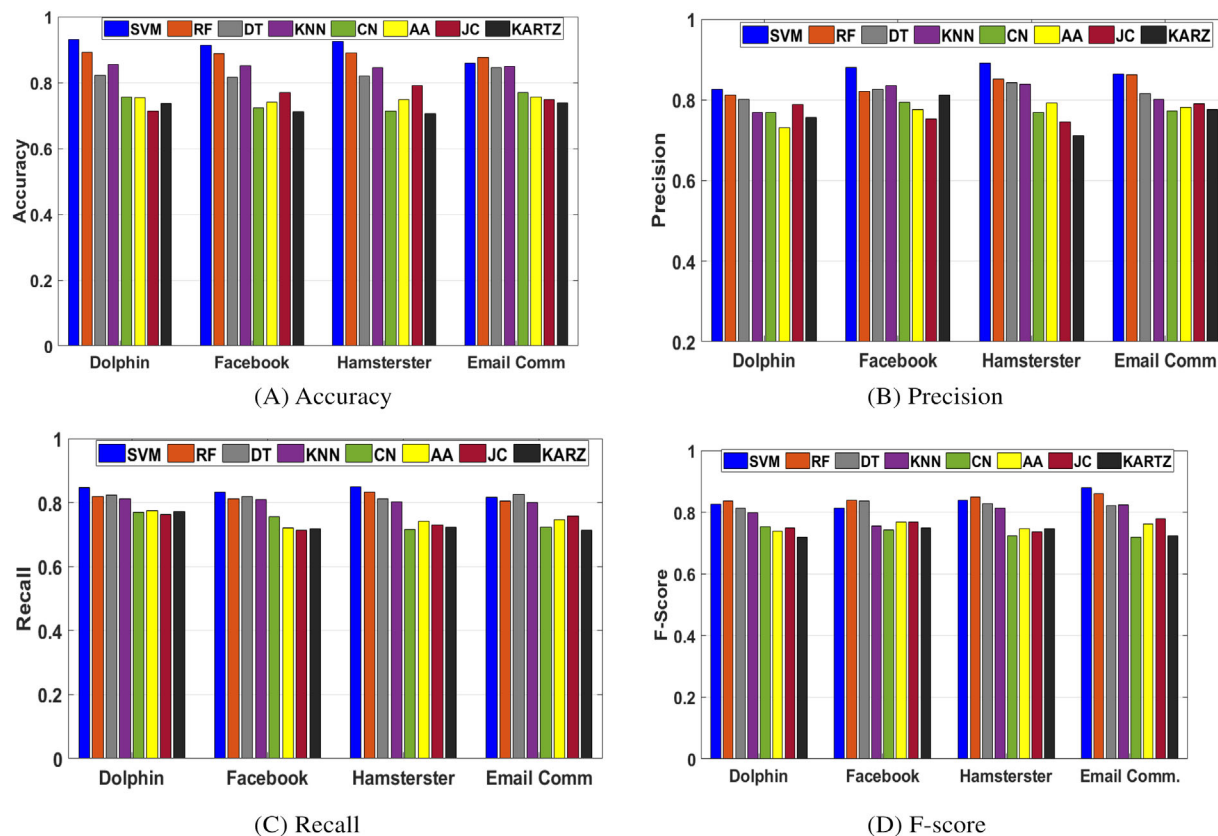


FIGURE 3 Comparative analysis of several link prediction algorithms in real-world data sets

We have also compared all the link prediction models in terms of F-score. Figure 3D shows the comparative analysis in term of F-score. It can be observed that Random Forest algorithms perform better in Dolphin, Facebook, and Hamsterster networks. However, the performance of SVM dominates all the algorithms in the email communication network. CN seems to have the worst performance on Facebook, Hamsterster, and email Communication network as it only considers the information around the neighborhood nodes.

From all the observations in Figure 3A-D, It can be observed that the performance of machine learning algorithms dominates all the structured-based algorithms like CN, JC, KARZ, and AA. SVM is found to have the best performance in terms of accuracy, precision, and recall. However, the Random Forest algorithm has better performance in term of F-score for all real-world networks.

7.4 | Results analysis of synthetic network data sets

Apart from the real-world networks, we have experimented on four artificial networks that are generated using LFR benchmark, as presented in Table 2. All the networks have been generated by fixing up the various parameters, as discussed in Table 3. The graphical comparison of various link prediction algorithms in synthetic network is presented in Figure 4A-D. It can be observed that SVM is found to have better performance in terms of precision, recall, accuracy, and F-score. KARZ algorithm has the worst performance in all the evaluation measures except in precision. It can be noted that all the machine learning algorithms have dominated the structured-based algorithms not only in real-world networks but also in synthetic networks.

The AUC value for different algorithms is presented in Table 4. It can be observed that the SVM gives better AUC value for all the considered data sets. The AUC value for machine learning models are observed to be high compared with the other structured-based link prediction algorithms. The boxplot analysis of performance in term of accuracy, precision, recall, and F-score is presented in Figure 5A-D, respectively. It can be observed that the mean performance of SVM is better in all the metrics.

After the performance analysis phase, we have ranked the set of nonexisting links based on the similarity score obtained from the best algorithm. In our case, SVM is found to be the best classifier model for link prediction compared with other algorithms. We have ranked the edges for real-world networks only. Table 5 lists out the 10 links that are currently not present in the network but are most likely to appear in the future.

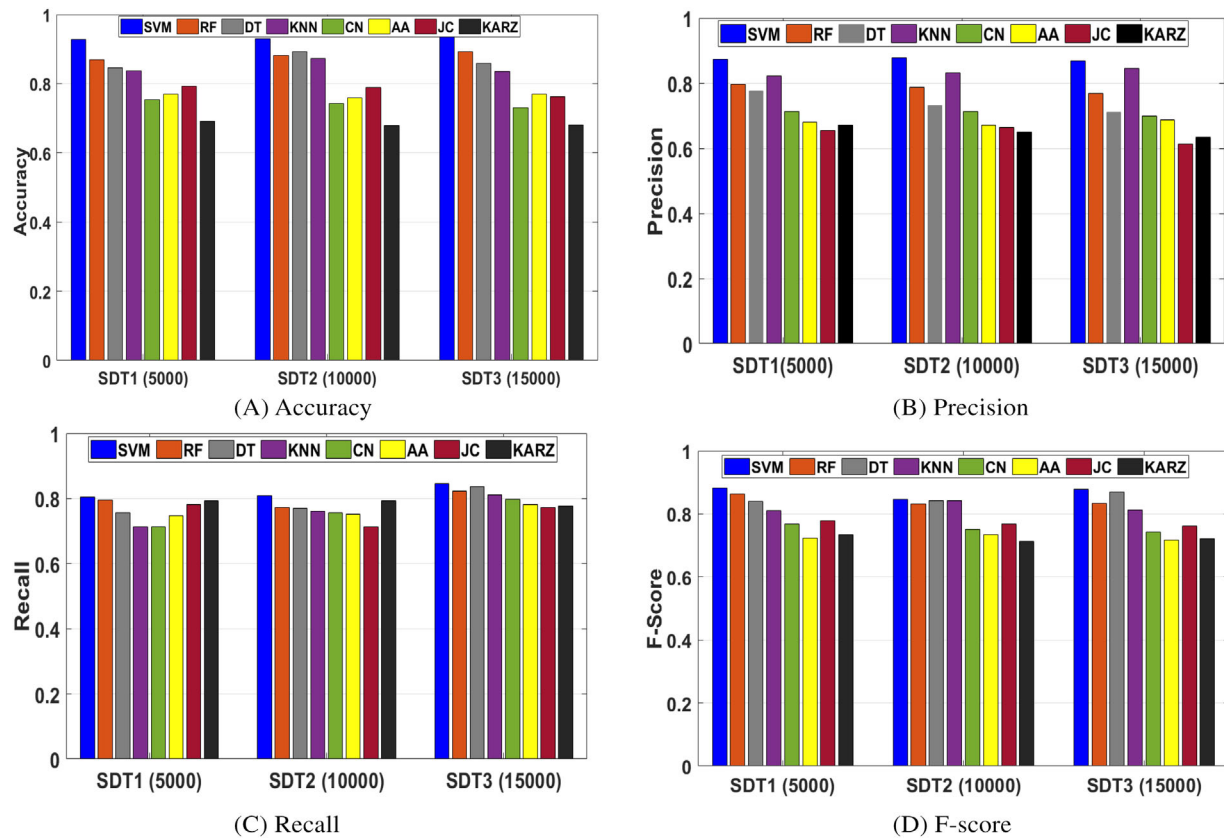


FIGURE 4 Comparative analysis of several link prediction algorithms in synthetic data sets

TABLE 4 Comparative analysis of link prediction algorithms in term of AUC value

SVM	RF	DT	KNN	CN	AA	JC	KARZ
0.901	0.821	0.811	0.769	0.622	0.639	0.746	0.745
0.912	0.882	0.805	0.812	0.654	0.672	0.754	0.776
0.891	0.865	0.812	0.805	0.795	0.803	0.811	0.826
0.923	0.865	0.855	0.849	0.836	0.723	0.736	0.796
0.916	0.886	0.876	0.866	0.842	0.821	0.832	0.798
0.908	0.899	0.881	0.879	0.811	0.786	0.781	0.791
0.936	0.882	0.856	0.844	0.799	0.759	0.779	0.783

8 | CONCLUSION AND FUTURE WORK

Link prediction in complex networks is an emerging research domain in social network analysis. It could be applied in modeling and understanding the evolution of social groups in a network. Such interpretation can help us for the effective implementation of models to discover hidden groups or the absent relationships in the groups. One of the popular applications of link prediction is to develop the recommending system, which is definitely helpful in modeling efficient business strategy. This is widely adopted in many real-world applications, especially in the e-commerce business. Another application of link prediction could be data analysis in the area of security and criminal investigation research. The potential threat can be identified in the terrorist network using these link prediction models.

In this article, we have suggested machine learning-based link prediction models, which have considered the structured-based features of social networks to predict the missing links. A number of structured-based similarity measures are identified in the literature, which are the potential approaches for link prediction. It has been observed from the experimental analysis that any individual similarity measure fails to predict the links accurately. However, when they are considered as the features in the machine learning algorithm, links are predicted more accurately. We have used the traditional evaluation parameters like F-measure, accuracy, precision, and recall to validate our work. From the

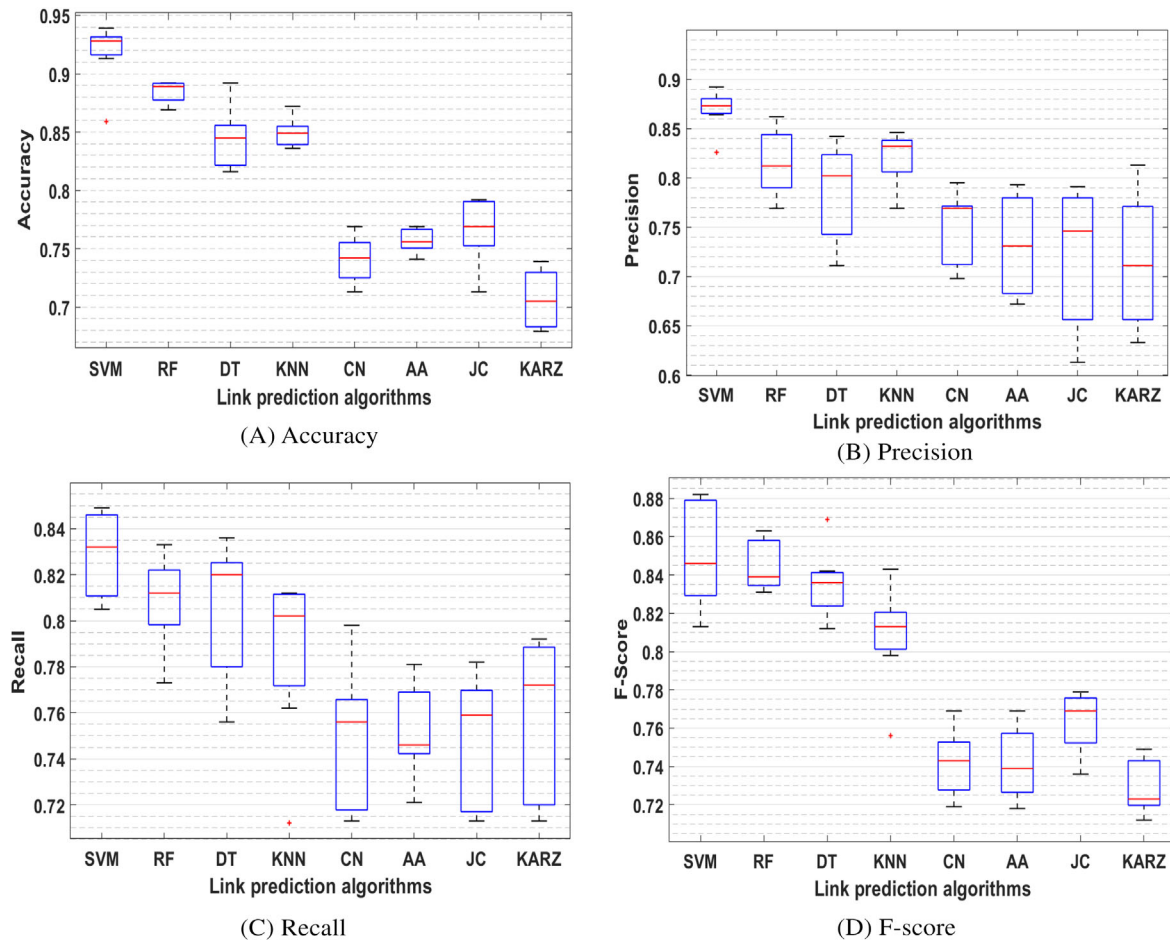


FIGURE 5 Box plot analysis of various link prediction algorithms

Rank	Dolphin	Hamsterster Friendship	Facebook	Email communication
1	(2,55)	(760,796)	(769,117)	(1755,288)
2	(16,29)	(796,1801)	(117,994)	(288,1731)
3	(14, 40)	(1776,79)	(769,88)	(1763,1743)
4	(38,9)	(762,787)	(135,997)	(1725,1730)
5	(37, 15)	(796,430)	(135,994)	(1717,288)
6	(18, 26)	(578,786)	(34,817)	(603,2761)
7	(21, 37)	(79,786)	(770,135)	(1525,2827)
8	(22, 39)	(1707,305)	(636,135)	(2783,603)
9	(6, 16)	(569,430)	(135,601)	(603,2687)
10	(5,47)	(777,79)	(722,135)	(1525,2707)

TABLE 5 Top 10 hidden link that may be established in future

observation from experiments, it is observed that the link prediction problem can be handled effectively by modeling it as a classification problem.

In our work, we have modeled the link prediction task as a classification problem in the area of machine learning, which has considered only the structural features to construct the feature vector. In the future, the content-based features could also be adopted for feature vector construction for accurate prediction. For the sake of simplicity, we have considered only the static and unweighted network to predict future links. In the future, the dynamic network may be considered, where links and nodes are added dynamically over time. In a weighted network, the weights associated with the links are one kind of measure to quantify the strength of bonding. This could also be a potential source for link prediction algorithms. This can be embedded in the proposed model for predicting future links.

ACKNOWLEDGMENTS

This research work was supported by Veer Surendra Sai University of Technology (VSSUT), Burla, Odisha, India. The authors wish to express their gratitude and heartiest thanks to the Department of Computer Science & Engineering, VSSUT, Burla, India for providing their research support.

ORCID

Kshira Sagar Sahoo  <https://orcid.org/0000-0002-6435-5738>

Anand Nayyar  <https://orcid.org/0000-0002-9821-6146>

Ashish Kumar Luhach  <https://orcid.org/0000-0001-8759-0290>

REFERENCES

1. BeheraRanjan Kumar, Naik Debadatta, Sahoo Bibhudatta, Rath Santanu Ku. *Proceedings of the 9th Annual ACM India Conference*. Centrality approach for community detection in large scale network. 2016; 115–124. <https://dl.acm.org/doi/10.1145/2998476.2998489>.
2. Behera Ranjan, Rath Santanu, Misra Sanjay, Damaševičius Robertas, Maskeliunas Rytis. Large Scale Community Detection Using a Small World Model. *Applied Sciences*. 2017;7(11):1173. <https://doi.org/10.3390/app7111173>.
3. Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. *J Am Soc Inf Sci Tech*. 2007;58(7):1019–1031.
4. Al Hasan M, Chaoji V, Salem S, Zaki M. Link prediction using supervised learning. Paper presented at: Proceedings of the 30 of SDM06 Workshop on Link Analysis, Counter-Terrorism and Security; 2006:798–805.
5. Behera Ranjan Kumar, Sahoo Kshira Sagar, Mahapatra Sambit, Rath Santanu Kumar, Sahoo Bibhudatta. Security issues in distributed computation for big data analytics. *Handbook of e-Business Security*. BocaRaton, FL: Auerbach Publications; 2018:167–190.
6. Lü L, Zhou T. Link prediction in complex networks: a survey. *Phys A Stat Mech Appl*. 2011;390(6):1150–1170.
7. Daminelli S, Thomas JM, Durán C, Cannistraci CV. Common neighbours and the local-community-paradigm for topological link prediction in bipartite networks. *New J Phys*. 2015;17(11):113037.
8. Fire M, Tenenboim L, Lesser O, Puzis R, Rokach L, Elovici Y. Link prediction in social networks using computationally efficient topological features. Paper presented at: Proceedings of the 2011 IEEE 3rd International Conference on Privacy, Security, Risk and Trust and 2011 IEEE 3rd International Conference on Social Computing; 2011:73–80; IEEE.
9. Chen H, Li X, Huang Z. Link prediction approach to collaborative filtering. Paper presented at: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'05). 2005: 141–142IEEE.
10. Murata T, Moriyasu S. Link prediction of social networks based on weighted proximity measures. Paper presented at: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence IEEE Computer Society; 2007:85–88.
11. Li RH, Yu JX, Liu J. Link prediction: the power of maximal entropy random walk. Paper presented at: Proceedings of the 20th ACM International Conference on Information and Knowledge Management; 2011: 1147–1156; ACM.
12. Lü L, Jin CH, Zhou T. Similarity index based on local paths for link prediction of complex networks. *Phys Rev E*. 2009;80(4):046122.
13. Meng B, Ke H, Yi T. Link prediction based on a semi-local similarity index. *Chin Phys B*. 2011;20(12):128902.
14. Papadimitriou A, Symeonidis P, Manolopoulos Y. Fast and accurate link prediction in social networking systems. *J Syst Softw*. 2012;85(9):2119–2132.
15. Dong L, Li Y, Yin H, Le H, Rui M. The algorithm of link prediction on social network. *Math Probl Eng*. 2013;2013.
16. Gao F, Musial K, Cooper C, Tsoka S. Link prediction methods and their accuracy for different social networks and network metrics. *Sci Program*. 2015;2015:1.
17. Sadeque F, Solorio T, Pedersen T, Shrestha P, Bethard S. Predicting continued participation in online health forums. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*. 2015;12–20.
18. Yao L, Wang L, Pan L, Yao K. Link prediction based on common-neighbors for dynamic social network. *Proc Comput Sci*. 2016;83:82–89.
19. Haghani S, Keyvanpour MR. A systemic analysis of link prediction in social network. *Artif Intell Rev*. 2019;52(3):1961–1995.
20. Zhou K, Michalak TP, Wanek M, Rahwan T, Vorobeychik Y. Attacking similarity-based link prediction in social networks. Paper presented at: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. International Foundation for Autonomous Agents and Multiagent Systems; 2019:305–313.
21. Newman ME, Moore C, Watts DJ. Mean-field solution of the small-world network model. *Phys Rev Lett*. 2000;84(14):3201.
22. Opsahl T, Vernet A, Alnuaimi T, George G. Revisiting the small-world phenomenon: efficiency variation and classification of small-world networks. *Organiz Res Methods*. 2017;20(1):149–173.
23. Abdul R, Paul A, Gul M, Hong WH, Seo H. Exploiting small world problems in a SIoT environment. *Energies*. 2018;11(8):2089.
24. Adamic LA, Huberman BA. Power-law distribution of the world wide web. *Science*. 2000;287(5461):2115–2115.
25. Zhou K, Michalak TP, Rahwan T, Wanek M, Vorobeychik Y. Adversarial link prediction in social networks. *CoRR*. 2018.
26. Wang S, Tang J, Aggarwal C, Chang Y, Liu H. Signed network embedding in social media. Paper presented at: Proceedings of the 2017 SIAM International Conference on Data Mining; 2017:327–335; SIAM.
27. Li X, Fang H, Zhang J. Rethinking the link prediction problem in signed social networks. Paper presented at: Proceedings of the 31st AAAI Conference on Artificial Intelligence; 2017.
28. Kumar Behera R, Kumar Rath S, Misra S, Damaševičius R, Maskeliunas R. Distributed centrality analysis of social network data using mapreduce. *Algorithms*. 2019;12(8):161.
29. Giridhar N, Bharadwaj K. Signed social networks: a survey. Paper presented at: Proceedings of the International Conference on Advances in Computing and Data Sciences; 2016:326–335; Springer.
30. Yuan W, He K, Guan D, Zhou L, Li C. Graph kernel based link prediction for signed social networks. *Inf Fusion*. 2019;46:1–10.
31. Yaghi RI, Faris H, Aljarah I, Ala'M AZ, Heidari AA, Mirjalili S. Link prediction using evolutionary neural network models. *Evolutionary Machine Learning Techniques*. New York, NY: Springer; 2020:85–111.
32. Alzubi JA. Optimal classifier ensemble design based on cooperative game theory. *Res J Appl Sci Eng Tech*. 2015;11(12):1336–1343.

33. Alzubi O, Alzubi J, Tedmori S, Rashaideh H, Almomani O. Consensus-based combining method for classifier ensembles. *Int Arab J Inf Tech (IAJIT)*. 2018;15(1).
34. Alzubi JA. Diversity based improved bagging algorithm. Paper presented at: Proceedings of the International Conference on Engineering & MIS; Vol 2015, 2015:1-5.
35. Srilatha P, Manjula R. Similarity index based link prediction algorithms in social networks: a survey. *J Telecommun Inf Tech*. 2016;2:87-94.
36. Ahmad A, Khan M, Paul A, et al. Toward modeling and optimization of features selection in big data based social Internet of Things. *Future Generat Comput Syst*. 2018;82:715-726.
37. Ahmad A, Babar M, Din S, et al. Socio-cyber network: the potential of cyber-physical system to define human behaviors using big data analytics. *Future Generat Comput Syst*. 2019;92:868-878.
38. Paul A, Ahmad A, Rathore MM, Jabbar S. Smartbuddy: defining human behaviors using big data analytics in social internet of things. *IEEE Wirel Commun*. 2016;23(5):68-74.
39. Sahoo KS, Tiwary M, Mishra P, Reddy SRS, Balusamy B, Gandomi AH. Improving end-users utility in software-defined wide area network systems. *IEEE Trans Netw Serv Manag*. 2019.
40. Rathore MM, Paul A, Hong WH, Seo H, Awan I, Saeed S. Exploiting IoT and big data analytics: defining smart digital city using real-time urban data. *Sustain Cities Soc*. 2018;40:600-610.
41. Din S, Paul A, Hong WH, Seo H. Constrained application for mobility management using embedded devices in the Internet of Things based urban planning in smart cities. *Sustain Cities Soc*. 2019;44:144-151.
42. Behera RK, Sukla AS, Mahapatra S, Rath SK, Sahoo B, Bhattacharya S. Map-reduce based link prediction for large scale social network; 2017.
43. Paul A, Jeyaraj R. Internet of Things: a primer. *Human Behav Emerg Technol*. 2019;1(1):37-47.
44. Adamic LA, Adar E. Friends and neighbors on the web. *Soc Netw*. 2003;25(3):211-230.
45. Lusseau D, Schneider K, Boisseau OJ, Haase P, Slooten E, Dawson SM. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behav Ecol Sociobiol*. 2003;54:396-405.
46. Hamsterster friendships network dataset – KONECT; 2017.
47. Facebook wall posts network dataset – KONECT; 2017.
48. Yin H, Benson AR, Leskovec J, Gleich DF. Local higher-order graph clustering. Paper presented at: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2017:555-564.

How to cite this article: Kumari A, Behera RK, Sahoo KS, Nayyar A, Kumar Luhach A, Prakash Sahoo S. Supervised link prediction using structured-based feature extraction in social network. *Concurrency Computat Pract Exper*. 2022;34:e5839. <https://doi.org/10.1002/cpe.5839>