Link Prediction Evaluation Using Palette Weisfeiler-Lehman Graph Labelling Algorithm

Salam Jayachitra Devi, Jawaharlal Nehru University, New Delhi, India
Buddha Singh, Jawaharlal Nehru University, New Delhi, India
Haider Raza, University of Essex, Colchester, UK

ABSTRACT

Link prediction is gaining interest in the community of machine learning due to its popularity in the applications such as in social networking and e-commerce. This paper aims to present the performance of link prediction using a set of predictive models. In link prediction modelling, feature extraction is a challenging issue and some simple heuristics such as common-neighbors and Katz index were commonly used. Here, palette weisfeiler-lehman graph labelling algorithms have been used, which has a few advantages such as it has order-preserving properties and provides better computational efficiency. Whereas, other feature extraction algorithms cannot preserve the order of the vertices in the subgraph, and also take more computational time. The features were extracted in two ways with the number of vertices in each subgraph, say K = 10 and K = 15. The extracted features were fitted to a range of classifiers. Further, the performance has been obtained on the basis of the area under the curve (AUC) measure. Comparative analysis of all the classifiers based on the AUC results has been presented to determine which predictive model provides better performance across all the networks. This leads to the conclusion that ADABoost, Bagging and Adaptive Logistic Regression performed well almost on all the network. Lastly, comparative analysis of 12 existing methods with three best predictive models has been done to show that link prediction with predictive models performs well across different kinds of networks.

KEYWORDS

Classifiers, Feature extraction, Graph labeling, Link prediction, Machine learning

DOI: 10.4018/IJKSS.2019010101
INTRODUCTION

Link prediction is outlined as a task of predicting whether any pair of nodes is likely to form links in the future by Lü and Zhou (2011). Recently, it has pulled in expanding consideration from both machine learning communities and data mining as given in Liben-Nowell and Kleinberg (2007). It has an extensive variety of applications in the field of recommendation systems, e-commerce, bibliography, bioinformatics, world wide web, etc. In an e-commerce domain, the most well-known usages of link prediction is in the construction of recommendation systems (Talasu et al., 2017). In the bibliography, link prediction is mainly used for record linkage and de-duplication according to Al Hasan et al. (2011). In case of bioinformatics, link prediction is used in predicting protein to protein interaction (Lü, & Zhou, 2011) (Xu, & Guan, 2014). Besides all the above applications, it has also been used in security purposed such as identifying a hidden group of radicals and criminal activities (Ozgul, Erdem, & Bowerman, 2009).

In the early stage, similarity scores obtained from different heuristic methods are used for predicting the links between the nodes in the network (Lü & Zhou, 2011), (Liben-Nowell, & Kleinberg, 2007). A heuristic score is assigned to each unconnected pair of nodes to measure the proximity of connection between the two nodes. Heuristic methods range from simple to more complex. Some of the simplest heuristics include common-neighbors (Liben-Nowell et al., 2007), Katz index (Katz, 1953), Adamic-Adar (Adamic et al., 2003), etc. Such heuristics work well in social network and is more scalable and interpretable. For heuristics, which lies in mid-level complexities include methods which used network topologies to calculate proximity scores. Later, many sophisticated and complex models have been developed, such as Matrix factorization (Salakhutdinov, & Mnih, 2008, July) and stochastic block model (Airoldi et al., 2008) which utilizes maximum likelihood estimation approaches.

However, in the last decade, an additional interest has arisen for researchers, to determine which method is most suitable for a specific situation. These heuristics are not universally suitable for different kinds of networks. For instance, the simplest heuristic common neighbor is more suitable in a social network for predicting friendships and it is also used in the collaboration network for prediction of co-authorship. This simplest heuristic has also a drawback, as it gives poor performance for networks such as biological networks and electrical networks (Lü & Zhou, 2011). On the contrary, the resistance distance (RD) gives a great performance in electrical and internet router networks (Klein & Randić, 1993). But the inverse happens in case of social networks. From the survey of over 20 different heuristics, it can be concluded that none of the heuristics perform well on all the different types of networks (Lü & Zhou, 2011). The existing heuristics used the topological features of the network. So, from the local patterns of each pair of nodes, one can automatically determine the possibility of formation of links. As the existing heuristics are not applicable across all the networks, the only resolution to find, which heuristics is suitable for a particular network is by using trial and error or expert selection (Brodhead, 2018).
Simultaneous Tolerance Synthesis for Manufacturing and Quality using Evolutionary Algorithms
[www.igi-global.com/article/simultaneous-tolerance-synthesis-manufacturing-quality/54344?camid=4v1a](www.igi-global.com/article/simultaneous-tolerance-synthesis-manufacturing-quality/54344?camid=4v1a)

Biological Traits in Artificial Self-Reproducing Systems
[www.igi-global.com/chapter/biological-traits-artificial-self-reproducing/28375?camid=4v1a](www.igi-global.com/chapter/biological-traits-artificial-self-reproducing/28375?camid=4v1a)

Early Stage Diagnosis of Eye Herpes (NAGIN) by Machine Learning and Image Processing Technique: Detection and Recognition of Eye Herpes (NAGIN) by Using CAD System Analysis