Supervised Learning for Link Prediction in Social Networks

Link prediction nelle reti sociali attraverso l'utilizzo di metodi e modelli di apprendimento supervisionato

Riccardo Giubilei, Pierpaolo Brutti

Abstract Link prediction is an estimation problem that has drawn a great deal of attention in recent years. In this work, a supervised learning approach is adopted to perform link prediction on data retrieved from Facebook. The specific goal, then, is to estimate the probability of two users to become friends in order to recommend them to one another whenever this probability turns out to be sufficiently high. On social platforms like Facebook, friendship recommendation is clearly a crucial ingredient since, when properly implemented, it plays a key role in determining the network growth. The contribution of this work consists in performing friendship recommendation on Facebook using a supervised learning approach that takes explicitly into account vertices' attributes; that is, all the personal information that users make available on their profiles.

Abstract La link prediction ha attirato molta attenzione negli ultimi anni. In questo lavoro, un approccio di apprendimento supervisionato viene utilizzato per fare link prediction su dati provenienti da Facebook. L'obiettivo è quindi quello di stimare la probabilità che due utenti diventino amici, in modo da suggerire gli uni agli altri quanto tale probabilità è alta. La raccomandazione delle amicizie è un problema molto importante poiché il suo corretto funzionamento è fondamentale per la crescita delle reti, che è l'obiettivo primario di siti come Facebook. Il contributo di questo lavoro è quello di fare ciò utilizzando un approccio di apprendimento supervisionato che prenda esplicitamente in considerazione anche gli attributi dei vertici, ovvero le informazioni personali che gli utenti inseriscono nel proprio profilo.

Key words: Link Prediction, Social Network Analysis, Network Science, Graph Theory, Supervised Learning, Machine Learning, Binary Classification.

Riccardo Giubilei
Sapienza University of Rome, e-mail: riccardo.giubilei@uniroma1.it

Pierpaolo Brutti
Sapienza University of Rome, e-mail: pierpaolo.brutti@uniroma1.it

1 Introduction and background

A social network is a popular way to model the interaction among people belonging to a group or a community. It can be represented using a graph, where each node is a person and each link indicates some form of association between two people.

In this framework, performing link prediction consists in predicting which nodes are likely to get connected. More precisely, the goal is to predict the likelihood of a future association between two unconnected nodes. This is carried out supposing the likelihood of link formation depends on the similarity between the two nodes.

In 2004, Liben-Nowell and Kleinberg [3] proposed one of the earliest link prediction models that works explicitly on social networks. The learning paradigm in this setup typically extracts the similarity between a pair of vertices exploiting various graph-based similarity metrics and uses the ranking on the similarity scores to predict the link between two vertices.

Subsequently, Hasan et al. [2] extended this work in two ways. First, they showed that using external data outside the scope of graph topology (namely, the vertices' attributes) can significantly improve the prediction results. Second, they used several similarity metrics as features in a supervised learning setup where the link prediction problem is posed as a binary classification task. Since then, the supervised classification approach has been popular in various other works on link prediction.

2 Motivation

The popularity of online social network services has thrived in recent years, attracting an increasingly large number of users. By connecting users with similar professional backgrounds or common interests, they open up new channels for information sharing and social networking. Creating connections not only helps to improve user experience, but also increases the chance of producing larger and more well-connected networks, which is the primary goal of these sites.

Consequently, link formation is fundamental. In Facebook, a link is formed whenever two people become friends. To increase the probability of link formation, users with the highest probability of becoming friends may be suggested to one another. This is achieved through the friendship recommendation system, that aims to find the most similar users in terms of their profile contents or their behavior so as to offer them to each other.

The scope of this analysis is to apply link prediction methods and techniques to perform friendship recommendation on Facebook data. The problem is tackled using a supervised learning approach that blends together both topological features and users' personal information. Incorporating the latter as covariates is everything but trivial so, in the following, we introduce a relatively simple method to handle effectively this crucial modelling step.

3 Data

The data was collected in 2014 by Julian McAuley (UC San Diego) and Jure Leskovec (Stanford University) using a Facebook application that asked to a pool of volunteers the permission to download their Facebook's profile information via Facebook API. In order to ensure the volunteers' privacy, all the data have been completely anonymized by assigning users and features sequential IDs. The data collection proceeded in the form of ego networks, i.e. starting from a central node (the person who gave the permission), and then expanding the network considering his friends and the mutual friends between them and the central node. 110 ego networks were collected, making a total of $27,520$ Facebook users. For each of these users, public information contained in their profile was also recorded.

Therefore, the dataset is composed by 110 distinct files, which correspond to the 110 ego networks, and by the additional file that contains the users' attributes. In this work we focus our attention on two specific ego networks. The first one is associated to user $6,934$ and has been selected as the train dataset being the largest among those that do not contain links to users from other ego networks. It is formed by 773 nodes, including the central one, and by $26,023$ links between them. Since the number of nodes is 773, the number of potential links in the network, given by all the possible combinations between the nodes, is $268,278$. Therefore, the number of actual links is approximately the 9.70% of all the possible links. The second one is the ego network of user $3,236$, and has been chosen as the test dataset for being structurally different from the first one. Indeed, it is composed of 345 nodes and $4,013$ links among them, which correspond to the 6.76% of the $59,340$ possible links.

4 Experimental setup

The link prediction problem is formalized as a supervised classification task, where each instance corresponds to a pair of vertices in the social network graph. Instances are characterized by features describing the similarity between the two nodes and a label denoting their link status. In particular, the instance is classified as positive if there exists a link between the nodes, or negative otherwise. The output of the models is a score for each non-observed link which quantifies how likely it is that it will actually become a link. The instances classified as positive are those that exceeds a certain threshold score.

Since each instance corresponds to a pair of vertices, the features should necessarily represent some form of proximity between them. In existing research works on link prediction, the vast majority of the features are related to the graph topology. Typically, they are built by computing similarities based on the node neighborhoods or on the set of paths that connect those two nodes. However, as anticipated in the Introduction, Hasan et al. [2] have proposed to extend the set of features in order to include also the vertices' attributes.

Now, coming back to our specific application, since we are dealing with ego networks, their topology immediately implies a diameter equal to 2. As a consequence, any feature based on paths is definitively not very informative and will not be included in the analysis. On the other hand, five neighborhood-based similarity indices that single out different aspects of the link formation phenomenon are selected. More specifically we consider: *Common Neighbors*, *Jaccard Index*, *Preferential Attachment Index*, *Adamic-Adar Index* and *Resource Allocation Index* [5]. For what concerns other *local* indices available in the literature, it is enough to say that they will not be considered here mainly because they have already shown to not lead to significant improvements in similar analyses.

Attributes-based features are built considering the file containing the users' attributes related to their personal Facebook profile. However, this file contains some redundant information, and, in addition, many of the attributes collected are not available for the majority of the users. Redundant attributes, such as the first, the middle, the full name and the ID, are excluded from the analysis. Likewise, all attributes that have been recorded for less than $1,000$ users are not considered. Among the remaining ones, some additional feature selection is carried out, eliminating variables with little to no informativeness. In order to build similarity indices from the remaining attributes, it is important to underline that a user may have more than a value for the same attribute. Consequently, the idea is to count, for each pair of nodes and for each attribute, the number of values they have in common for that attribute. This is motivated by the belief that the larger the number of characteristics two unconnected users have in common, the higher the probability that they will be linked in the future. This procedure leads then to a data-matrix, with the rows corresponding to the pair of vertices, and the columns being the attributes. The generic entry for this matrix is the number of times the values of a certain attribute coincide for the pair of nodes considered.

5 Models and results

Five binary classification models are considered: *Random Forest*, *Neural Network*, *Gradient Boosting*, *Naive Bayes* and *Logistic Regression*. For each model, a careful parameter tuning is carried out. In order to evaluate the predictive abilities of these models, a 10-fold cross validation is performed on the train data. The models are then evaluated using a number of metrics, including *accuracy*, *specificity*, *recall*, *precision*, *F1 score*, *Area Under the Receiver Operating Characteristic curve* (AUROC) and *Area Under the Precision-Recall Curve* (AUPRC).

Table 1 shows the performance comparison for the different classifiers considered. For the fixed-threshold metrics, the threshold has been set to 0.5. The results are very good for almost every metric. However, the choice of the best model is performed by considering only the metrics that are independent of the threshold chosen to convert the probability scores to class labels, i.e. AUROC and AUPRC.

Consequently, the best model is the Gradient Boosting, which is then used to make prediction on the test data.

Model	Accuracy	Specificity	Recall	Precision	F1 score	AUROC	AUPRC
Random Forest	89.97%	89.48%	94.57%	49.11%	64.65%	97.24%	79.43%
Neural Network	94.37%	97.53%	64.88%	73.87%	69.09%	97.11%	78.74%
Gradient Boosting	94.49%	97.64%	65.15%	74.81%	69.45%	97.24%	79.53%
Naive Bayes	92.31%	93.29%	83.26%	57.15%	67.77%	96.44%	74.86%
Logistic Regression	91.07%	91.01%	91.69%	52.27%	66.58%	96.99%	76.72%

Table 1: Evaluation metrics for the models considered using a 10-fold cross validation.

6 Prediction

The results obtained using the Gradient Boosting model on the test data are reported in Table 2. In addition to the threshold-independent metrics AUROC and AUPRC, also recall and precision are included, being of interesting and useful interpretation in the specific context. In fact, a recall of 82.63% indicates by definition that a little more than 8 people out of 10 a user may want to add are indeed suggested. On the other hand, a precision equal to 70.17% means that users would add approximately 7 people out of 10 suggested.

Model	Recall	Precision	AUROC	AUPRC
Gradient Boosting	82.63%	70.17%	98.06%	84.66%

Table 2: Evaluation metrics for the prediction on the test data.

Figure 1 allows to visualize the predictive results obtained using the model. In particular, all the effectively existing links are reported in the figure, coloring them

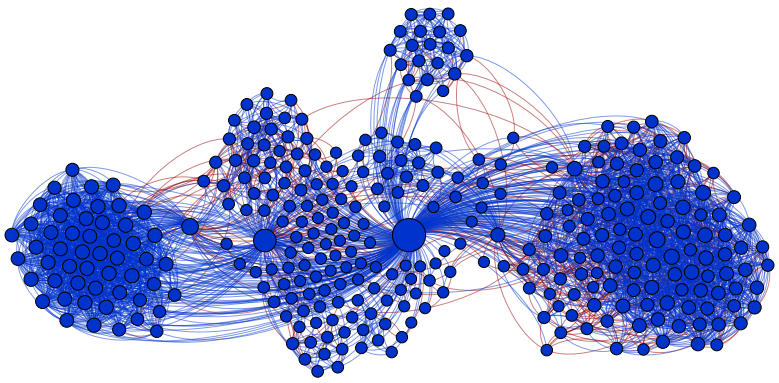

Fig. 1: Correctly predicted links (blue) and missed links (red) in the test ego network.

differently based on the predictions: correctly predicted link are colored in blue and missed links are colored in red. Therefore, it is possible to note that the model manages to reconstruct a massive part of the graph. This is achieved at the cost of suggesting only 3 people that a user would not be interested in adding every 10.

7 Conclusions and future work

The results obtained show that the link prediction task has been accomplished in a satisfying way. In particular, the inclusion of attribute-based features seems to be extremely useful, allowing to make very good predictions. This aspect is of great importance for both confirming that including them actually improves the predictive abilities of the models, and for validating the procedure used to build them starting from the users' personal information. In addition, the supervised learning approach has proved to be effective also for performing link prediction on Facebook data, with particular reference to ego networks.

In the future, it would be interesting to consider larger graphs, and study specific methods and techniques that scale well on "big data" networks.

In addition, the possibility to retrieve and consider the time-stamps of the links is certainly an aspect which would help in link prediction. For example, they may be included by treating more recent links as more important than older ones. Important contributions on this extension, also defined time-aware link prediction, are those by Ahmed et al. [1] and by Tylenda et al. [4].

The techniques and models presented here may be exploited to perform link prediction outside the specific task of recommending friends on Facebook. For example, the case of directed networks may be considered, being of great importance in other online social networks like Twitter. In the end, link prediction is a very relevant problem in almost every kind of networks, and it would be interesting to consider applications also in these other domains.

References

1. Ahmed, A., Xing, E. P.: Recovering time-varying network of dependencies in Social and biological studies. In: Proceedings of the National Academy of Sciences of the United States of America, 106(29): 11878–11883 (2009)
2. Hasan, M. A., Chaoji, V., Salem, S., Zaki, M.: Link Prediction using Supervised Learning. In: Proceedings of SDM Workshop of Link Analysis, Counterterrorism and Security (2006)
3. Liben-Nowell, D., Kleinberg, J.: The Link Prediction Problem for Social Networks. Journal of the American Society for Information Science and Technology, 58(7): 1019–1031 (2004)
4. Tylenda, T., Angelova, R., and Bahadur, S.: Towards Time-aware Link Prediction in Evolving Social Network. In: SNA-KDD '09: Proceedings of the third Workshop on Social Network Mining and Analysis (2009)
5. Wang, P., Xu, B., Wu, Y., Zhou, X.: Link Prediction in Social Networks: the State-of-the-Art. Science China Information Sciences, 58(1): 1–38 (2005)