SML Project1: Who are my friends?

Xiaowen JIN, Zhizhang LIN, Wei LI

1 Introduction

Link prediction is a problem of increasing importance in the digital world, which is predicting the potential existence of a link between a pair of users based on the current state of relationships in the social network. In this project, we use Twitter data to train algorithms and to classify a real edge between two nodes from fake. Different similarity methods are used to calculate the probability of the edge exists between nodes. For each method, we get the common similarity, the similarity of outbound neighbors and inbound neighbors prospectively. Furthermore, machine learning models are introduced using similarity as features. The evaluation results indicate the best way is inbound Cosine similarity.

2 Data

The training data is crawled from Twitter social network in the form of adjacency lists with nodes as user IDs. The training network is constructed by 20,000 users noted as *Source* and their following lists, the user IDs in the following lists are noted as *Sink*. The following relationship between a *Source* and a *Sink* is called a *link*. There are 4,867,136 user IDs in this network and 24,004,361 links.

A sub-train set of size 20,000 including 10,000 positive pairs and 10,000 negative pairs is extracted from the training data because of two problems: the original training data contains only positive labels, indicating that there is a link between a pair (*Source*, *Sink*); and the original training data is too large to process. 200 *Sources* and their corresponding *Sink* lists are randomly sampled from training data, then all the existing links are assigned with positive labels, and other pairs without a link are assigned with negative labels. A validation set of size 5,000, with 2,500 positive pairs and 2,500 negative pairs is extracted from the training data in the same way as the sub-train set.

3 Related Work

In social networks, link prediction mines missing links in current network, which benefits the friendship recommendation system (Aiello et al., 2012), social tie prediction (Liu, Li, & Wang, 2020), and authorship identification (Ozcan & Oguducu, 2019). The taxonomy of the link prediction technologies consists of three bottoms: similarity, learning methods, and social theory (Wang, Xu, Wu, & Zhou, 2014; Kumar, Singh, Singh, & Biswas, 2020). Combining similarity and learning methods is a fashion in link prediction, Xie, Gong, Wang, Liu, and Yu (2019) proposed an embedding method based on similarity and obtained a promising prediction accuracy, and Manshad, Meybodi, and Salajegheh (2020) merged similarity and time series. These works inspired our team to feed similarity features to machine learning methods.

4 Feature

In this paper, similarity measurements are served as the features to describe the links in the Twitter network. This idea is motivated by the intuition that for a pair of nodes with higher similarity, it is more likely to potentially form a link between the pair. The similarity between node *a* n_a and node *b* n_b can be divided into three types (Cukierski, Hamner, & Yang, 2011):

Type I: $Score(n_a, n_b) = similarity(n_a, n_b)$

Type II : $Score(n_a, n_b) = \frac{1}{|\Gamma_{out}(n_a)|} \sum_{n \in \Gamma_{out}(n_a)} similarity(n, n_b)$

Type III : $Score(n_a, n_b) = \frac{1}{|\Gamma_{in}(n_b)|} \sum_{n \in \Gamma_{in}(n_b)} similarity(n_a, n)$

$\Gamma_{out}(n_a)$ means the outbound set of a, $\Gamma_{in}(n_a)$ denotes the inbound set of a, and $\Gamma(n_a)$ indicates the union of the outbound set and the inbound set of a, for example, in the network that a follows b and c and d follows a, then $\Gamma_{out}(n_a)$ is $\{b, c\}$, $\Gamma_{in}(n_a)$ is $\{d\}$, and $\Gamma(n_a)$ is $\{b, c, d\}$.

It is noteworthy that Type II contains much more information than Type III in the social network, since users can control who they want to follow, but can not control who wants to follow them. Further, all of the three are based on one assumption, human can be depicted by their social relationships. And the similarity functions *similarity*(n_a, n_b) are neighbor-based similarity shown in Table 1.

Table 1: DESCRIPTION OF THE SIMILARITY

Name	Formula	Description
Adar (Adamic & Adar, 2003)	$\sum_{n\in\{\Gamma(n_a)\cap\Gamma(n_b)\}} \frac{1}{log\lvert\Gamma(n)\rvert}$	Weights connections with rare nodes more heavily
Common Neighbors (CN)	$\lvert\Gamma(n_a)\cap\Gamma(n_b)\rvert$	Num. common neighbors
Cosine	$\lvert\Gamma(n_a)\cap\Gamma(n_b)\rvert/(\lvert\Gamma(n_a)\rvert\lvert\Gamma(n_b)\rvert)$	Common neighbors divided by preferential attachment (Newman, 2001)
Jaccard	$\lvert\Gamma(n_a)\cap\Gamma(n_b)\rvert/(\lvert\Gamma(n_a)\rvert\cup\lvert\Gamma(n_b)\rvert)$	Common neighbors divided by total neighbors

Table 2: The AUC OF SIMILARITY AND MODEL

Name	Type I	Type II	Type III
Adar	**0.80140**[a]	—[b]	0.64677
CN	0.73067	0.68997	0.62689
Cosine	0.79743	0.69522	**0.90504**
Jaccard	0.73706	0.78049	**0.87331**

Name	AUC
RF	**0.85922**
LR	0.79229

[a] The AUC score greater than 0.8 is bold.

[b] Computing Type 2 of Adar costs too much time, thus we decide to ignore it.

5 Model

By treating this link prediction task as a binary classification problem, two machine learning algorithms random forest and logistic regression are used to train a classification model, which returns not only labels, but also a probability of the existence of a potential link between a pair of users.

5.1 Random Forest

A random forest model (RF)[1] is made up of a large number of uncorrelated decision trees, then combines the predictions made by those decision trees to produce a more accurate prediction (Ali, Khan, Ahmad, & Maqsood, 2012).

5.2 Logistic Regression

The second model used in this paper is logistic regression (LR)[2]. It is a classification algorithm. It is used to predict a binary outcome based on a set of independent variables. Its output is between 0 and 1 which is the probability of the two nodes' links in this project (Hosmer Jr, Lemeshow, & Sturdivant, 2013).

5.3 Model training

It is assumed that as more features are used to describe a link, the model will be better trained. Hence all different similarity scores of each link in the sub-train data set are calculated and used as input features to train both random forest and logistic regression models. Then a hyper-parameter tuning with 5-fold cross validation[3] on both models are performed.

6 Evaluation

A receiver operating characteristic curve (ROC curve) is a probability curve to illustrate the performance of a binary classifier with various discrimination threshold (Fawcett, 2006). The area under curve (AUC) of ROC measures model separability. A model is considered to have a good separability if its AUC is close to 1, otherwise close to 0 (Myerson, Green, & Warusawitharana, 2001).

In this project, we use AUC to evaluate the performance of the methods, and the results are displayed in Table 2. The Type III Cosine similarity outperforms among all the similarity methods, and achieves an AUC score of 0.90504 for the test data, and Random Forest outperforms logistic regression model and achieves an AUC score of 0.85922.

[1] Implemented by Sklearn Package in Python: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[2] Implemented by Sklearn Package in Python: https://scikit-learn.org/stable/modules/generated/sklearn.linear$_{m}$*odel.LogisticRegression.html*

[3] Implemented by Sklearn Package in Python: https://scikit-learn.org/stable/modules/generated/sklearn.model$_{s}$*election.GridSearchCV.html*

7 Discussion

7.1 Data

Training data could contain more information about the users such as Tweets and the personal information, thus the similarity should not only include graph information among neighbors and links, but the user's characteristics. Secondly, the negative data constructed for model training may cover some potential positive links.

Further, some insignificant links are produced by the nodes with a large inbound/outbound set. For example, the official account of BBC in Twitter has an inbound set with 171 million elements, those links hardly represent the *Source's* characteristic and the similarity between the users.

7.2 Feature

The similarity used as features only considers the first degree neighbors[1] of the users due to the limited computation ability, while higher degrees of neighborhood contains more information about the graph. The path information in the graph between two users is that given a pair of *Source* and *Sink*, whether it is possible for *Source* to reach *Sink* through paths in the graph. This information can not be represented by neighbor-based similarity, and is missing from the features. The reason for type III Cosine outperforming Type II is that the data has more inbound information than outbound, since the number of *Sink* is much greater than that of *Source*. Though Type I of Cosine merges the outbound and inbound information, this imbalance combination deteriorate the performance. According to the definition of Cosine similarity, Cosine is more sensitive in directed data (graph), thus it outperforms than other similarity measurements. In further development, weighting the outbound and inbound information can obtain a more balance similarity.

7.3 Model

The AUC score for random forest is higher than that for logistic regression; and the similarity methods generally performed better than model methods. In social network, users tend to follow the similar sinks, which triggers the correlations among links. This dependency is conflict with the observation independence assumption of logistic regression, thus LR does not perform well. Further, random forest is trained merely by the neighbor-based similarity features, these homogeneous features degenerates the performance of RF (Tang, Garreau, & von Luxburg, 2018). The weak links in the data also cause the decision boundary is unclear between positive and negative labels.

8 Conclusion

Despite the flourish of the algorithms and frameworks in computer science, link prediction in big data is still a demanding problem. Due to the time and computation limitation, some promising methods, such as Node2vec and Artificial Neural Network, were not tested. Our simple method performs well in this particular data set, and it is unlikely to handle the material world. However, this method can provide a solid foundation for future complex models.

It is common to treat social network as a graph, while the edges can represent various types of connections between nodes, and it is difficult for nodes to contain text or picture information. Our similarity methods achieved the best performance over models, illustrating the importance of expressing the information directly.

References

Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. *Social networks*, *25*(3), 211–230.

Aiello, L. M., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., & Menczer, F. (2012). Friendship prediction and homophily in social media. *ACM Transactions on the Web (TWEB)*, *6*(2), 1–33.

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, *9*(5), 272.

Cukierski, W., Hamner, B., & Yang, B. (2011). Graph-based features for supervised link prediction. In *The 2011 international joint conference on neural networks* (p. 1237-1244).

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*(8), 861–874.

[1] 1-degree neighbor (a, b): a -> b; 2-degree neighbor (a, c): a -> b -> c; 3-degree neighbor (a, d): a -> b -> c -> d; and so on.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Kumar, A., Singh, S. S., Singh, K., & Biswas, B. (2020). Link prediction techniques, applications, and performance: A survey. *Physica A: Statistical Mechanics and its Applications*, *553*, 124289.

Liu, Z., Li, H., & Wang, C. (2020). New: A generic learning model for tie strength prediction in networks. *Neurocomputing*.

Manshad, M. K., Meybodi, M. R., & Salajegheh, A. (2020). A new irregular cellular learning automata-based evolutionary computation for time series link prediction in social networks. *Applied Intelligence*, 1–14.

Myerson, J., Green, L., & Warusawitharana, M. (2001). Area under the curve as a measure of discounting. *Journal of the experimental analysis of behavior*, *76*(2), 235–243.

Newman, M. E. (2001). Clustering and preferential attachment in growing networks. *Physical review E*, *64*(2), 025102.

Ozcan, A., & Oguducu, S. G. (2019). Multivariate time series link prediction for evolving heterogeneous network. *International Journal of Information Technology & Decision Making*, *18*(01), 241–286.

Tang, C., Garreau, D., & von Luxburg, U. (2018). When do random forests fail? In *Advances in neural information processing systems* (pp. 2983–2993).

Wang, P., Xu, B., Wu, Y., & Zhou, X. (2014). Link prediction in social networks: the state-of-the-art. *CoRR*, *abs/1411.5118*. Retrieved from `http://arxiv.org/abs/1411.5118`

Xie, Y., Gong, M., Wang, S., Liu, W., & Yu, B. (2019). Sim2vec: Node similarity preserving network embedding. *Information Sciences*, *495*, 37–51.