Predicting co-authorship in academic co-authorship networks

XinTong Ma 1028582

1 Introduction

Modelling the interaction among the community through the social network has attracted much attention in recent years. One of the popular research direction is predicting prospective links in co-authorship network. This is known as link prediction in a complex network, which refers to predicting the probability of a connection between two nodes in a graph. This paper investigated the collaborations between two scientists and predict latent connections. The co-authorship graph is formed by nodes which correspond to authors and the edges which represent the collaborations. The links were either labelled 0 or 1 depending on whether the co-authorship exists. Features were extracted based on the network topology and node attributes. Several supervised learning methodology was applied to construct binary classifier models.

2 Related Work

Machine Learning methodology such as decision tree and Support Vector Machine (SVM) was widely used in link prediction. Al Hasan, Chaoji, Salem, and Zaki (2006) stated that SVM with radial basis function (RBF) kernel outperforms all other algorithms with a robust margin. The squared error of SVM is 30% less than any others. This is also agree by Pavlov and Ichise (2007). However, they used a linear kernel with sequential minimal optimization training algorithm to model the data. One of the challenge in link prediction problem is the skew of data (Doppa, Yu, Tadepalli, & Getoor, 2010). For thousands of nodes, the amount of possible link is uncountable. The number of actual links is a tiny fraction of total amount. The skew in positive and negative class leads to the difficulty of model training. Al Hasan and Zaki (2011) suggested that altering the samples size of training sets could handle with class skew. Thus, the model require more positive-label data than negative class especially in SVM which is norm-based (Al Hasan et al., 2006).

3 Methodology

The modelling consists of four components, fake edge construction, feature engineering and predictor selection and evaluation. As mention above, a lack of actual link has severe damage to the prediction. Thus, the number of negative edges was down-sampling to 10000 and the number of positive edges remains 26000.

Features have a direct impact on the performance of classifiers. According to Al Hasan et al. (2006), proximity features such as keyword match count, aggregated features such as a sum of keyword count, and topological features such as clustering index are popular features in link prediction problem. The features based on the identity of authors are justified to outperforms more (Wang, Xu, Wu, & Zhou, 2015). The fact that the amount of primary neighbours is dominant in link prediction is well-known. Al Hasan et al. (2006) proposed that the number of secondary neighbours sometimes plays an important role. Since the number of secondary neighbours grows exponentially, the logarithm of a secondary degree was taken. Features such as Jaccard Coefficient, Clustering Index etc. were widely used in link prediction (Al Hasan et al., 2006). These features are related to graph characteristics and they were justified to be a benefit in graph problem.

In addition, the raw binary features regarding the keywords and venues of authors' publication are sparse and high dimensional. Principal Component Analysis (PCA) was used to reduce the dimension of binary features. The impatiences of all feature were assessed by computing the variance and correlation.

Three predictors were used to classify the links: logistic classifier, SVM and XGbooster. A logistic classifier is widely used in binary classification. It is simple to implement. SVM with rbf kernel

was implemented using Scikit Learn and XG boost was implemented by XGBoost package from Github (Chen & Guestrin, 2016; Pedregosa et al., 2011). The parameters selection of these models were achieved via grid search in Scikit Learn package and cross-validation so that the optimal solution could be obtained. In this experiment, performances of classifiers were accessed by the Area Under Curve (AUC). Compare with evaluation metrics such as precision-recall curves, ROC curves decouple effectiveness of predictor from error cost and class skew (Yu et al., 2014). Therefore, the quality of probabilistic classifier can be measured by AUC.

4 Result and Evaluation

The features with top 10 correlation which listed in Figure 1 do not provide optimal solution. Features such as clustering index with lowest have a great contribution to the models. The explanation of this result is, correlation can only indicate linear relationships between features and labels. Variance threshold is neither a suitable feature selector. It is highly affected by the scale of the feature value. A small scale but valuable feature threatens to be filtered out through variance threshold. After feature selection using PCA with estimators equal to 2, the training dataset is clearly separable as shown in Figure 2. However, the outcome of PCA on the test set is not ideal. Figure 3 indicated that the decision boundary between classes is still blurry after using feature processed by PCA.

Not all features provide significant enhancement on the prediction performance. The features that found to be valuable are based on the identity between nodes and topological structure but not aggregated features. Both proximity and topological features enhancing the model with around 25% AUC score. For those features that are found to be significant, the characteristics of the distribution of feature value are generally noticeable. For example, edges that are predicted to be fake have less than 3 common neighbours. The larger the amounts of common neighbour and common secondary neighbour are, the higher chance that nodes will be linked. For topological feature such as Adamic Adar Index, Jaccard Coefficient, if the value is high then the probability of linkage is high. However, the probability does not decrease even though the value of the feature is small. Some researchers stated that non-topological features are more advanced in improving the performance of prediction but this promising effect is not shown in this dataset (Wang et al., 2015). Topological features top ranked features as shown in Table 1.

The AUC score of three classifier is approximately similar as shown in Table 2. It can not be concluded that which classifier offer a best prediction. The outcome of prediction is not ideal. The difference in distributions of train and test dataset could be one of the cause. Among all classifiers, XGboost classifier can handle the sparse feature and it is more powerful in preventing over fitting as it provides regularisation. However, XGboost is weak in processing high dimension data. It is the best approach for dataset with several features.

5 Conclusion

This research aims to predict the likelihood of future connections between vertices. Link in a real-life network is proven to be predictable. A down-sampling was conducted to handle the dataset imbalance. Topological features and proximity features are beneficial enough in link prediction. Moreover, logistic classification, SVM, and XGboost provide solutions with acceptable accuracy. XGboost is recognized as the best approach as it have ability to deal with over fitting and sparse features.

Feature	Aggregated	Proximity	Topological
AUC	0.655	0.882	0.923

Table 1: Performance of features

Classifier	Baseline	Logistic	SVM	XGBoost
AUC	0.675	0.932	0.921	0.932

Table 2: Performance of Classifiers

References

Al Hasan, M., Chaoji, V., Salem, S., & Zaki, M. (2006). Link prediction using supervised learning. In *Sdm06: workshop on link analysis, counter-terrorism and security* (Vol. 30, pp. 798–805).

Al Hasan, M., & Zaki, M. J. (2011). A survey of link prediction in social networks. In *Social network data analytics* (pp. 243–275). Springer.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/2939672.2939785` doi: 10.1145/2939672.2939785

Doppa, J. R., Yu, J., Tadepalli, P., & Getoor, L. (2010). Learning algorithms for link prediction based on chance constraints. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 344–360).

Pavlov, M., & Ichise, R. (2007). Finding experts by link prediction in co-authorship networks. *FEWS*, *290*, 42–55.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Wang, P., Xu, B., Wu, Y., & Zhou, X. (2015). Link prediction in social networks: the state-of-the-art. *Science China Information Sciences*, *58*(1), 1–38.

Yu, Q., Long, C., Lv, Y., Shao, H., He, P., & Duan, Z. (2014). Predicting co-author relationship in medical co-authorship networks. *PloS one*, *9*(7).

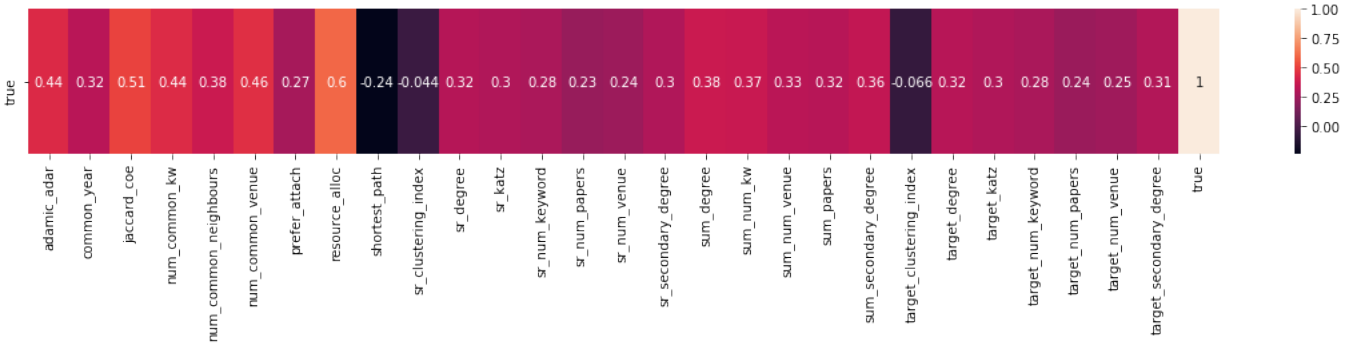

Figure 1: Correlation between features and label

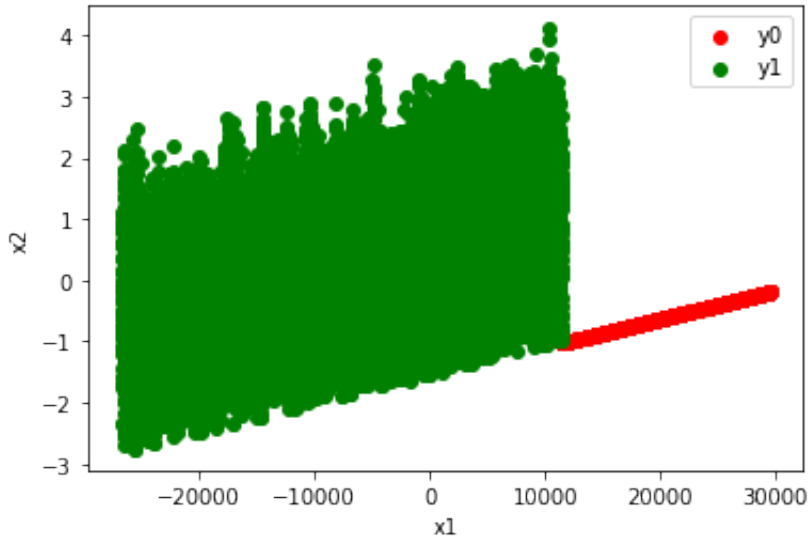

Figure 2: Visualisation of PCA effect on train set

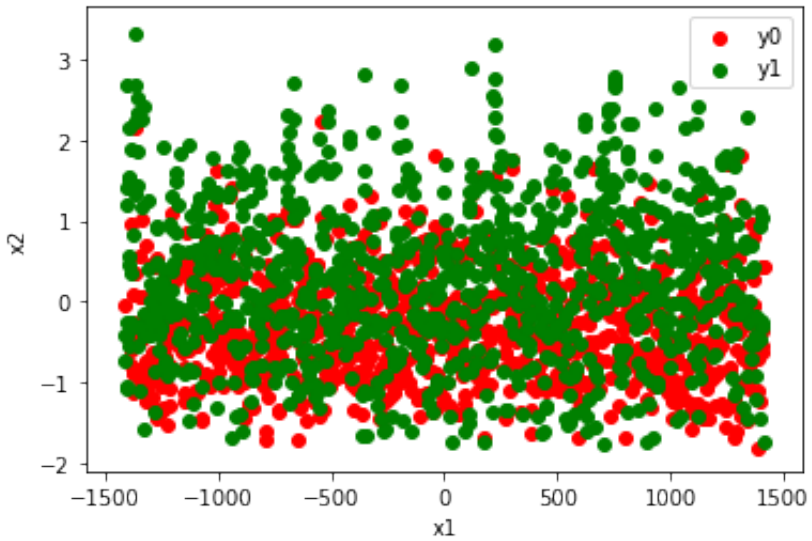

Figure 3: Visualisation of PCA effect on test set