COMP90051_2020_SM2 Statistical Machine Learning Project

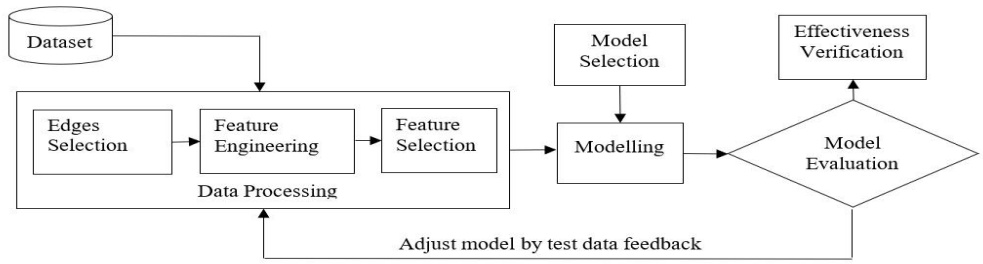
Group 87 Xiaolin Zhang, Zihan Ye, Yilin Yu

1. Introduction

Through the known social network relationship within a period of time, we can predict the future interaction between members. We call this problem Link Prediction Problem. It has other common applications, such as friend recommendation on social networking sites, predicting the interaction between proteins, predicting the relationship between suspects, product recommendations, etc. The report is to introduce the processing of the link prediction through the data analysing, feature and model selection, advantages and disadvantages discussion. Finally, it is about our reflection and how to improve in the future.

2. Experiments and Dataset Description

The flow chart of this project is displayed in Graph 1.In this project, The training data contains 20000 records crawled from Twitter which own 4867136 nodes. The test data contains 2000 edges, 1000 of which are real and the other 1000 do not truly exist. Our task is to train the data to distinguish true edge from false edge so that to predict whether edges exist among test node pairs in the test data.



Graph 1. Experimental Pipeline

2.1 Sample Approach

For sample generation, we created an array ‘source’ which contains all source nodes and an array ‘sink’ which contains all sink nodes for sampling. The size of two arrays are 20000 and 4867136.

For positive samples, we firstly choose a node ‘a’ from ‘source’ randomly, and then choose a node that sourced from ‘a’ randomly. The node pair is considered as a positive sample. We repeat the process until we get enough samples.

For negative samples, we firstly choose a node from ‘source’ randomly, and then choose a node from ‘sink’ randomly. If the source node is not the source of the sink node, we consider the node pair as a negative sample, otherwise we do the process again. We repeat the process until we get enough samples.

2.2 Feature Selection

In this section, we survey an array of methods for link prediction. The training data given can easily be converted to a graph. We use library ‘networkx’ to construct the undirected graph. And use the adjacency matrix to represent the directed graph. In graph theory, there are various indexes that can be used to measure the properties of nodes, which can help us to generate the features of edges. For each pair of nodes, we extract features based on the basic property of nodes and some conventional algorithms.

We would make a brief description of the basic algorithms used to generate features. All the algorithms assign a connection weight score(x,y) to pairs of nodes <x,y>. The description of five undirected features is shown in Table 1, and the description of two directed features is shown in Table 2.

Jaccard's coefficient (JC)	Used to compare the similarities and differences between a limited sample set. The larger the Jaccard coefficient value, the higher the sample similarity.
Adamic Adar (AA)	An intimacy measurement method based on common neighbours between nodes.
Resource Allocation (RA)	Similar to *Adamic Adar*, except that the logarithmic penalty is removed, but it works better in many networks.
Preferential Attachment (PA)	Depends on the number of connections between two nodes. Larger number of connections between two nodes indicates greater the probability of labelling the node pair being connected.
Sum of Degree (DE)	Sum of the degree of two nodes.. In a social network, a larger value of DE represents more connections with other users.

Table 1. Description of undirected features

Predecessor Similarity for Source and Sink (PS)	Value the similarity between the predecessor vector of two nodes
Successors Similarity for Source and Sink (SS)	Value the similarity between the successor vector of two nodes

Table 2. Description of directed features

To analyze which feature is the most important in our model and select the best feature combination, we generated our own testset to test and evaluate our model. We removed some features, retrained the model, and reevaluated the model. The results are listed in Table 3. The results reveal that: a) The model built based on both directed features and undirected features would perform better than only use one type of features. b) The model built based on undirected features performs better than directed features, which indicates indirected features are important in measuring the relationships. c) Different undirected features may have similar effects on the performance of the model.

Finally we selected all features for classification. While the twitter network is a directed graph, these features reflect edge properties in both the directed graph and undirected graph.

Features	AUC	Features	AUC
<JC,AA,RA,PA,DE,PS,SS>	*0.828*	*<JC,AA,RA,DE,PS,SS>*	*0.824*
<JC,AA,RA,PA,DE>	*0.789*	*<AA,RA,PA,DE,PS,SS>*	*0.824*
<PS,SS>	*0.726*	*<JC,RA,PA,DE,PS,SS>*	*0.824*
<JC,AA,RA,PA,PS,SS>	*0.824*	*<JC,AA,PA,DE,PS,SS>*	*0.824*

Table 3. Results based on different features

2.3 Model Selection

Labeled data with extracted features should be plugin into the classifier algorithm to train a model to classify the edge into linked or not linked. Also, this is a binary classification problem. The algorithm selection is

based on binary classification problems. Then the data missing problem is considered. In the train set, for source nodes, its target nodes are listed integrally. Considering our sampling method, whatever positive edges or negative edges originate from a known source node, therefore the label of edge is integrated. The properties of an object are based on node features in the graph. Not all nodes are contained in the actual network generated from the train set, but neighbors of source nodes mostly remained. Good data integrity is an indicator of algorithm selection.

Logistic regression is applicable to binary classification problems. Also, its model shows good interpretability to show different features' impact on result. SVM is also a good classifier to solve the binary classification problem. SVM can train a model with a small train set. SVM is significant to compare with other classifiers to check whether the train set should be expanded. Random forest is also selected to deal with complex features and potential data missing possibilities.

We build models based on different classifiers and the results of the models are shown in Table 4. Due to the large size of our training data, SVM doesn't perform well. The result of the Random Forest is much better than Logistic Regression. Finally we predicted the test data using Random Forest Classifier.

	Logistic Regression	Random Forest	SVM
AUC	0.77	0.82	0.50

Table 4. Performance of different classifiers

3.Advantages and Disadvantages Analysis

We get all features based on the whole graph. The complete neighbor information of the source node can increase the performance of edge features and avoid the impact of data missing.

We select samples randomly from the whole dataset. We just selected 20000 samples for training, which is just a little part of the whole training data. The sampling method will generate a subgraph with high depth but difficult to generate a subgraph to reflect the features of the whole data set.

JC, AA, RA, PA and DE are all the features based on the neighbor nodes of the source node and target node of an edge. The number of neighbors is considered but the detailed properties of neighbor nodes are not involved. Thus the model based on only undirected features doesn't perform well enough.

PS and SS features are involved into the feature set. PS and SS consider the direction relation between the node and their neighbor. This is a directed graph so the involving of directed features is significant. But actually, the two features cannot reflect the properties of the edges comprehensively. More features can be involved such as clustering features. But these features are limited by the size of the whole network with 4,867,136 nodes which will cost a lot of computing resources.

4 Conclusion

In this report, we adopted the Random Forest algorithm with 7 features to reach the result. Through the selection of different models and features, we have made certain progress in the combination of directed and undirected features. For the future, we can continue to learn more about the prediction of directed features. And for the model, we will also improve it by learning new research methods.

5. References

[1] Ajay Kumar, Shashank Singh, Kuldeep Singh, Bhaskar Biswas(2020). *Link Prediction Techniques, applications and performance: A survey*. Department of Computer Science and Engineering, Indian Institute of Technology(BHU), Varanasi, 221-005, India