Author Cooperation Prediction

Student account: ziyliu2

1. Introduction

This report is an explanation of how to use machine learning to predict whether given two authors will cooperate. In this experiment, we first extracted all the author-pairs (positive edges) of cooperating authors from the training data, then randomly generated the same number of negative edges and combine with positive edges, and finally formed the final generated training-used training set by adding labels and a certain number of features. Through a certain number of attempts, 12 features are added and Random Forest Algorithm is used for this project. The prediction accuracy of the final model reached 90.8%. This report mainly discusses the selection of features and classifier algorithms and whether the selected features and algorithms have achieved the desired effect.

2. Generation of Training Set and Testing Set

2.1 Components of Dataset

The training set of an item is composed of the names, features and labels of each element in the training matrix. The name set of the training set elements of this project consists of all author pairs (positive edges) and the same number of uncooperative author pairs (negative edges), where the positive edge has a label of 1 and the negative edge has a label of 0. The selection of element characteristics will be described in detail below:

- Common neighbors: larger the number of common neighbors, larger the possibility that two authors know each other, larger the possibility that two authors cooperate.
- Common keywords: larger the number of common keywords, larger the possibility that two authors study in same field, larger the possibility that two authors cooperate.
- Common venues: larger the number of common venues, larger the possibility that two authors cooperate.
- Jaccard similarity[1]: the jaccard coefficient is defined as the ratio of the size of the intersection of two sets A and B to the size of its union:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

  In this project, neighbor, keyword and venue sets are used to calculate their Jaccard similarity separately. Larger Jaccard similarity, larger the possibility that two authors cooperate since their study field are more similar.
- Preferential Attachment[1]: product of two degrees (neighbors). If the two authors have collaborated with more people separately, the more likely they are to know and collaborate.
- Adamic-Adar[1]: formula of the algorithm is shown below:

$$A(X,Y) = \sum_{Z \in (k(X) \cap k(Y))} \frac{1}{log|k(Z)|}$$

  Taking the inverse of the logarithm is equivalent to calculating a weight for each common neighbor. The greater the degree of a common neighbor, the lower the weight. In other words, if there are too many people who have worked with a common neighbor, then the probability of introducing two people will be relatively low.

- Cosine similarity: Each element in the training set can form a vector. The smaller the cosine of the angle between the two vectors (a and b), the higher the similarity between the two elements.

$$\cos(\theta) = \frac{a \times b}{|a||b|}$$

  In this project, keyword cosine similarity and venue cosine similarity are used.
- Time difference: there are two features that need to be considered: (1) the greater the time difference between the publication of the latest paper by the two authors, the less likely it is to collaborate; (2) the longer the time overlap of two authors' paper publishing period, the more likely they collaborate together.

After making the dataset, it is divided to two part: 80% of the dataset is used as training set, another part is testing set to assess model accuracy.

3. Classifier

3.1 Classification Method

Two classifier are used in this project – Decision Tree and Random Forest: (1) the grouping of data in the Decision Tree is based on the values of attributes of the given data. The data items are split according to the values of these features. This process is recursively applied to each split subset of the data items and terminates as for as all the data items in current subset belong to the same class; (2) Random Forest is the forest formed by several random trees, which are constructed randomly from a set of possible trees with some random features on each node.

3.2 Classifier Performance

After testing, Random Forest Algorithm performs better than Decision Tree. Reason is shown below [2]:

- The number of features is large, the decision tree uses all the features and samples, which is prone to overfitting, random forest There are many decision trees constructed using part of the features of some samples (replacement sampling is adopted). Features and data are reduced on a single decision tree, reducing the possibility of overfitting.
- Unlike Decision Tree Algorithm, random forest selects part of the data to build multiple decision trees, even if the prediction on a single tree will be inaccurate due to the influence of outliers, the prediction results referring to multiple decision trees will reduce the impact of outliers.

4. Conclusion

In the project period, the best prediction score (90.8%) is made by using 12 features and Random Forest Algorithm. However, it does not perform well in competition, and still have a lot to improve:

- Dataset generation: try to generate unconnected node pairs of different hops in proportion (for example, 40% of 2 hops, 60% of random pairs).
- Analysis of features: draw a plot for a single feature (e.g. keyword) may help how to fit the data.
- Cross-validation can be used to use all the data in the data set to train and test the model to improve accuracy.

Reference

[1] L. Adamic, E. Adar. Friends and neighbors on the web. Soc. Networks, 25(3), 2003.
[2] Ali, Jehad & Khan, Rehanullah & Ahmad, Nasir & Maqsood, Imran. (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues(IJCSI). 9.