Academic Collaboration Prediction in JISE, ISECON, and ISEDJ

Procedure

1. Combine JISE, ISECON, and ISEDJ data, scraped from respective websites
2. Seperate data by publication date into 3 bins: 2000-2004, 2005-2009, 2010-2014
3. Form a coauthorship graph from publications in each bin
4. Form training and cross-validation sets:
   - Training Set: pairs formed between bin 1 and bin 2 (positive) and randomly sampled authors that did not collaborate (negative)
   - Cross Validation Set: pairs formed between bin 2 and bin 3 (positive) and randomly sampled authors that did not collaborate
5. Compute below features for each sample
6. Normalize data so each feature has a mean of 0 and std. of 1
7. Fit Logistic Regression, Support Vector Machine, and Random Forest Model on Training set
8. Measure accuracy on cross-validation set.

Features

Common Neighbors

$$|N(u) \cap N(v)|$$

Where u and v are both nodes, and $N(v)$ denotes the set the of all neighbors of node v. Common Neighbors was chosen as a feature under the assumption that two authors that share a large number of co-authors may have a higher chance of working together in the future.

Jaccard Coefficient

The Jaccard Coefficient measures similarity between two sets by dividing the size of the intersection by the size of the union:

$$\frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$$

Two nodes that have a high Jaccard Coefficient have very similar neighbors, which might be a good indication for future collaboration.

Resource Allocation Algorithm

Introduced in 2009 by Tao Zhou, Linyuan Lu, and Yi-Cheng Zhang in Predicting Missing Links via Local Information, the Resource Allocation algorithm is defined as:

$$\sum_{w \in N(u) \cap N(v)} \frac{1}{|N(w)|}$$

The idea behind the resource allocation algorithm is that if many of the common neighbors between u and v have a low number of neighbors themselves, any "resources" sent from u have a high likelihood of making their way to v and vice versa.

Preferential Attachment

Preferential Attachment is simply the product of the size of each node's neighbor set:

$$|N(u)||N(v)|$$

Two nodes that both have high numbers of neighbors, regardless of their commonality, may have a greater chance of collaboration in the future.

Results

Method	Accuracy
Logistic Regression	60.9%
SVM	60.9%
Random Forest	**67.4%**

With a random forest, we can use Mean Decrease in Impurity (MDI) to see what features are most conducive in predicting future collaboration:

Feature	Normalized MDI
Common Neighbors	0.04
Jaccard Coefficient	0.14
Preferential Attachment	**0.57**
Resource Allocation	0.25