Group 11:
Akshay Mehra
Sayam Ganguly
Patrick Green

November 7, 2016

Predicting Influencers in a Social Network

Objective of the Project:

To predict the human judgement about which of two users is more Influential in a social network, given twitter activity data for the two individuals. We model this as a Supervised Binary Classification problem. This problem initially featured as a Kaggle competition[1] in 2013.

Data:

Below are the numeric features that are provided for each user (in a pairwise format)

- Follower/Following Count
- Mentions Sent/Received
- Retweets Sent/Received
- Post Count
- Listed Count – a count of user-created lists or categories in which the user is included
- "Network Features" – proprietary, calculated features which describe the local follower network

A training set of 5500 user pairs is provided with a binary result indicating which user is evaluated to be more influential. Another 5900+ data pairs are provided as part a test set.

Methods:

We will utilize each of the following supervised learning algorithms for Binary Classification and compare the performance:

- Logistic Regression
- Neural Network
- Support Vector Machines
- Decision Trees

Additionally, we will explore the effectiveness of the following pre-processing methods with each algorithm:

- Feature Transformations (log, binary, other), using the difference of feature values
- Feature Selection Analysis, using algorithms such as Select-K-Best in scikit-learn library.

A submission mechanism is provided on Kaggle to evaluate predicted results for the test set, which measures the area under the Receiver Operating Characteristic (ROC) curve. Each model will be evaluated using the submission result on the test set, and by splitting the training data set to determine additional metrics (Area Under Curve, precision, accuracy, recall, etc.)

Tools:

- Microsoft Azure machine learning studio
- Python: scikit-learn

Team Roles:

There will be overlap in the contributions of each user to validate results and ensure that all members have full understanding of each piece of project. The iterative testing of specific algorithms may be divided amongst team members.

[1]https://www.kaggle.com/c/predict-who-is-more-influential-in-a-social-network