Criss-Crossing the Org Chart: Predicting Colleague Interactions with R

Eric Sun[1,*]

1. Facebook, 1601 S California Ave. Palo Alto, CA 94304

*Contact author: esun@facebook.com

Keywords: Machine Learning, Organizational Behavior

In this talk, we present a novel application of R for machine learning in the workplace: automatically predicting future levels of interaction between co-workers.

In mid-to-large sized organizations such as Facebook, employees typically work on projects with colleagues across many teams. In a fast-paced work environment, it quickly becomes difficult to remember the colleagues an employee has worked with and to identify the co-workers with whom she is most likely to interact with in the future. Such a system would have many useful applications:

- Suggesting peer reviewers during performance review season
- Optimizing seating charts for maximum productivity
- Setting up optimally-constructed teams within a company
- Automatically filtering internal feeds of employee content (such as commit logs) to deliver personalized content to each employee
- Suggesting new colleague interactions (based on second-degree connections) that may be useful to one's work
- Giving managers more insight into their employees' interactions

To accomplish this task, we generate a dataset where each row consists of a pair of employees. For each pair we calculate many features in several different categories: direct communication (such as the number of code reviews requested from one member of the pair to the other), implicit interaction (such as the number of meetings co-attended), and implicit communication (such as the number of common mailing list threads). As controls, we also include dummy variables for manager, direct report, and peer relationships, and also control for physical proximity (from seating charts). For each event, we weight each interaction by 1 / (number of people involved). Thus, a meeting with 5 people would count less than a one-on-one meeting.

Using these features, we create a model using the **randomForest** package in R that predicts the total number of weighted interactions between a pair of employees in the next 28 days using data from the previous 28 days. New predictions are generated automatically every night and are displayed in a dashboard for each employee to view. With the current model, mean square error on a held-out test set is 0.1049, and anecdotally, most employees have reported that predictions from the system are very accurate.

While the model is set up to predict future interactions, it is also interesting to examine the coefficients of the features to find out which of the independent variables leads to long-term, persistent interactions.

In addition to presenting results from the random forest model, we compare the performance and accuracy of various other algorithms including boosted trees (using the **gbm** package) and ordinary linear regression.