Predicting Doctor Drug Adoption with Networks

Jake Robbins 301208633
Department of Mathematics, Simon Fraser University

Objective

To acquire an understanding of why some doctors began prescribing a certain drug to their patients earlier than other doctors did:

- Use regression methods on available variables to form a predictive model.
- Analyse the network of relationships between the Doctors to aid our understanding.
- Integrate our network analysis with our predictive model to improve upon it.

Introduction

Our data consists of 242 doctors, each with recorded responses to 13 variables and connected to one another in a network by 4 kinds of relationships; 'friend', 'discussion', 'advice' 'random'. Figure 1 shows a plot of this network.
The 13 variables concern information about the doctors practice and professional life. The data is sourced from Coleman, Katz, & Menzel [1]. The variable that we shall build our model to predict is that of the adoption date of the drug tetracyclin by each doctor.

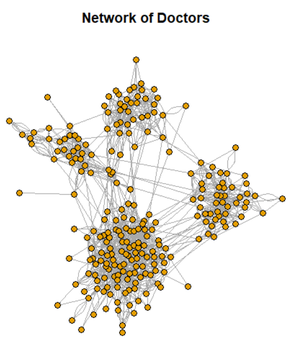

Figure 1: Relationships between doctors

Regression Model

We build a linear regression model to predict the 'Adoption Date' variable using the 12 other variables.

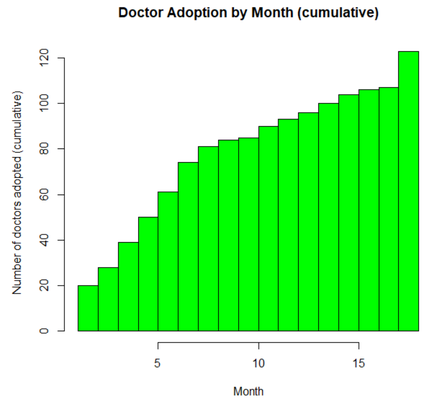

Figure 2: Adoption of the drug grew steadily in the first six months then began to slowly taper off in the next 12 months except for a large growth in the 18th month.

We randomly select 170 of our 242 nodes (approx. 70%) to use as training data for our linear model.
We then test our model on the remaining nodes using a 5-fold cross validation method.
With each fold we calculate the mean error over our test nodes and then take the mean of all 5 folds to receive a total mean error of:

2.3285714.

Centrality

Because there may be a social element that influences the adoption date variable, it is worthwhile to explore if the number or type of connections possessed by the doctors in our data affects our model.
Our network consists of 4 kinds of edges between nodes:'friend','discussion"network' & 'random'.
We use these to produce the following measures of centrality: Closeness, Betweenness, PageRank.

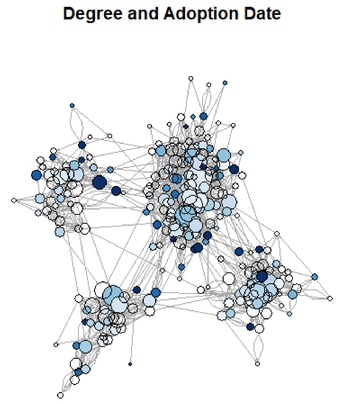

Figure 3: Larger nodes imply greater degree. Darker nodes imply a later adoption date.

We additionally create for each node a measure of degree for the 3 non-random edge types and an overall degree count. Figure 3 shows nodes adjusted by colour to indicate adoption date and adjusted by size to indicate degree.
We create a new linear model, this time including our centrality measures to total 19 predictors.

Missing Data

Across the 13 variables there was a significant amount of missing data. All variables were similarly affected so eliminating the most affected variables was not a viable option. There were several doctors who had no response to many of the 13 variables. As there were only 242 doctors included in our data, imputation of the missing values using a K-Nearest-Neighbours method was performed in favour of exclusion for these nodes.

Results

With our 19 predictors, including centrality measures, we construct a new linear model and once again do a 5-fold cross validation.
Our prediction model values all of the centrality measures as contributing to an earlier adoption date, with degree having the strongest correlation.
We take the mean of all 5 of the errors and our total mean error is:

1.6818048

Which is 27.77% smaller in error than our previous model that did not take the network into account.

Conclusion

Doctors with greater centrality measures in the Network were more likely to adopt the drug earlier in general than their less central peers. This may be because they are more likely be informed of or recommended the drug than others.

References

[1] Elihu Katz James Coleman and Herbert Menzel. The Diffusion of an Innovation Among Physicians. *American Sociological Association*, 20(4):253–270, Dec 1957.