CS229: Grant Prediction with Network Features

Stefan Krawczyk
Computer Science Department, Stanford University
stefank@cs.stanford.edu

ABSTRACT

This project investigates supervised learning to predict grant proposal approval or rejection, using the Mimir project's data on external grants applied for by Stanford faculty. Baseline features relating to the grant itself and faculty features are used, in particular the novel use of a faculty member's *network* to extract features is investigated and its effects on grant prediction.

1. INTRODUCTION

Grants take time and effort to put together and there is often anxiety related to their outcome. Wouldn't it be great for faculty members to assess their grant proposal before sending it out?

Lots of research lately has been done to try to study and analyze networks were people are represented as nodes and edges represent some type of relationship. In social networks this is a friend edge, in citation networks this is a cite from one paper to an older paper, or in collaboration networks the edge represents a link between co-authors. The properties of such networks then have been studied.

Burt[2] showed that people who sit in structural holes, or places of the network where they connect distant components, look to gain quicker promotions, better salaries, better ideas than their peers in the electronics company he analyzed. Burt[1] also produces *networks constraints* which help to detect these structural holes. This shows that there are certain characteristics of a network that could be amenable to a machine learning algorithm.

Having access to a Stanford faculty data set, ripe for feature extraction, containing information relating faculty to each other through relationship links, in addition to supplemental information about age, tenure, etc, a network of faculty as vertices and edges as relationships can be produced.

The saying "*success breeds success*"can also be hypothesized to mean that the relationships that one has with other people helps in determining your own success. Can we given the different network relationship information, help predict grant approval and rejection of Stanford faculty?

This paper is structured in the following way: section 2 gives an overview of the data, section 3 details the approach, section 4 talks about the feature sets, section **??** showcases the results and section 6 wraps up with the conclusion and future work.

2. DATASET

The data comprises of a collection of SQL databases, that was put together for the Mimir project wanting to study the flow of knowledge/ideas in an academic network. It contains information from 1995 to 2007 on:

- external grants applied for by faculty: the applicants; the amount proposed and awarded; sponsoring organization; school and department the grant is under; grant approval or rejection; and whether there are continuing grants.
- dissertation committees faculty sit on and thesis details of the student
- co-taught courses (from 1999)
- co-authored publications (not complete)
- supplemental information on the faculty like tenure status, age, gender, tenure status, position, number of appointments, courtesy appointments.

This data is both quite interesting and unique, as well as plentiful. The grant data in particular has over 22 thousand grant proposals, with a roughly 50/50 split in approvals and rejections.

In addition to grant prediction, there are many other options with which one could pursue both supervised or unsupervised learning. For instance the data is quite amenable to clustering; one could try to figure out which department courtesy appointments really belong to based on the data and see how that changes with time. There is also lots of prediction based tasks possible for example predicting departure of faculty, or number of students graduating in a given year, etc.

3. APPROACH

The data resided in SQL tables which after some manipulation was then output to CSV files. Feature extraction was then performed on the data to produce the three different feature sets.

Once the feature sets were produced they were fed into three different machine learning algorithms: logistic regression, a linear classifier, and a support vector machine (SVM). Combinations of the feature sets were explored to discern the impact of the different feature sets using two approaches:

1. randomly intermingling the years, not using any cumulative features, thus training and testing on grants from the entire time span. This was the quickest test but not a realistic test.
2. training on all previous years and testing on the next year - thus testing on each year from 1995 to 2007. This would be the realistic test as people would want to know how the current year's grant proposals are predicted.

Logistic regression was first used to test out the different feature sets before running the other classifiers, as the turn around time was much quicker.

The linear classifier was chosen because it was a readily available package and because the data was rather high dimensional, it was hypothesized that this classifier should be able to find some decent hyperplane. Initially gradient descent was used, but it kept getting stuck and was changed in favour to use the quasi-newton method.

Lastly the SVM with a linear kernel was picked because training it took the longest amount of time, and it was hoped that it should be as good if not better than the linear classifier.

3.1 Software Packages

There were two software packages that were used, the Java Universal Network/Graph[3] (JUNG) and the JavaNLP library.

JUNG was used for creating the network graphs so that feature extraction on the graphs could be performed. It was chosen because it was in Java and would interface well with the JavaNLP library and that it had a few network measures/metrics that would be of use in feature extraction.

The JavaNLP library was utilized for all of the machine learning algorithms.

4. FEATURES

The bulk of this project's work was in feature creation and extraction. In addition to what is described below, the log was taken of all continuous valued features, while the square was take only of all continuous valued features that would not cause an possible underflow or overflow to add to the feature mix. The following describes the three feature sets produced.

4.1 Grant Signature Features

These features were extracted entirely from the grant proposal itself. These were to be the simple baseline features that would be added to by the other feature sets.

It was comprised of: a bernoulli bag of word model on the project title, a bernoulli bag of word model on the sponsoring organization, the amount proposed, the school id the grant is from turned into a categorical feature, the deparment id the grant is from turned into a categorical feature, and the year.

4.2 Non-network Faculty Features

This feature set was based on anything else that could be extracted from the SQL tables that was not a network feature, that could be related to the faculty on the grant application. It was decomposed into three subsets: professor, department and school sets. When appropriate, cumulative features that spanned the history until the current year were also produced.

The professor set dealt with features directly attributable to one particular facutly member and was the largest set out of the three.

This set comprised of: gender, tenure status, primary department id (categorical feature), school id (categorical feature), primary job class code (categorical feature), number of appointments, H-index (was only available for a select number of faculty), number of courtesy appointments, number of dissertation students, number of publications (not complete for any faculty), ethnicity as a categorical feature, age, hire year, time at Stanford, total number of grants awarded last year, total number of grants rejected last year, total grants applied for last year, cumulative number of grants applied for, previous year's grant success rate, overall grant success up to last year, total amount proposed, total amount awarded, average amount proposed, ratio of total amount awarded over total amount proposed, average amount proposed for awarded grants, average amount proposed for rejected grants, and ratio of average proposed amount awarded over average proposed amount rejected.

The school and department sets dealt with aggregate features that covered the school and department respectively. The idea would be that would help cover and school or department wide attributes related to grants. These sets had features comprised of the aggregate grant related features from the professor set.

4.3 Network Faculty Features

This feature set was largest and took the most time to produce out of all the sets. It was produced by running a lot of network related metrics on the produced network graphs for each faculty member in the graph.

Eight graphs were produced per year: dissertation committees; (not complete) co-authorship; co-taught classes; new grants awarded in that year; grants continuing; grants continuing and awarded; grants rejected that year; and a combined edge set. The edges also contained weights representing the number of interactions that the faculty had together,

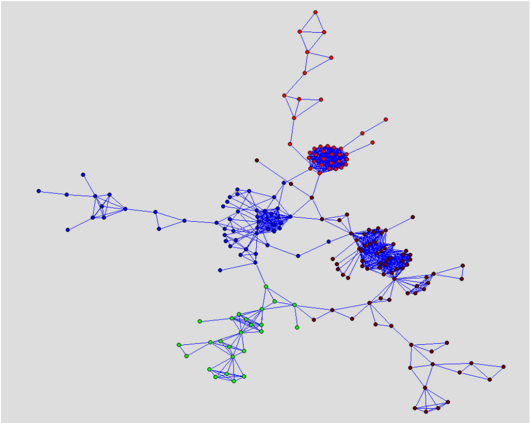

Figure 1: An example of a network graph created.

for feature extraction weighted as well as unweighted versions were produced where possible. An example graph can be seen in figure **??**

The following was extracted from each graph for each faculty member: five structural hole signature measures[1]: aggregate constraint, constraint, effective size, efficiency, hierarchy; barycenter value[1];random walk betweeness value[2]; betweeness centrality[3]; closenes centrality[4]; eigenvector centrality[5]; clustering coefficient[6]; number of total edges; radius one and two features: number of tenured and untenured links, number of awarded and continuing grants, number of rejected grants, number of grant and publication edges that overlapped, number of publications, neighbour edge incidence; and cumulative weighted, unweigted and average weight edge incidence.

5. RESULTS

For all the results we had a majority baseline of roughly 50% for each test.

5.1 Randomly Intermingling the years

5.1.1 Logistic Regression

In figure 2 we see the results from running logistic regression. The best testing accuracy came from the combined grant and network feature sets, so a positive result for the use of network features. From the graph we can also see that grant features by themselves do not give anything above baseline, while adding the other feature sets allows everything to be just above it.

The convergence parameters where fiddled with extensively for logistic regression, but did not yield any improvement over these results.

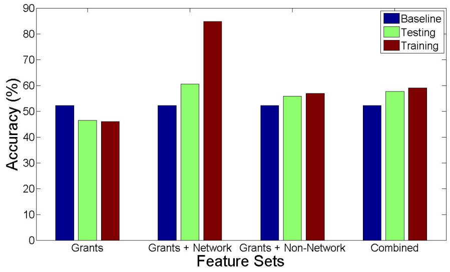

Figure 2: Results of testing on randomly intermingled years for logistic regression.

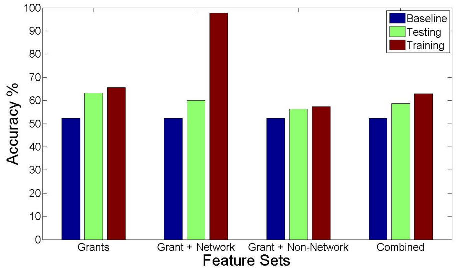

Figure 3: Results of testing on randomly intermingled years for the linear classifier.

5.1.2 Linear Classifier

Counter-intuitively to what the results were in figure 2, the grant features by themselves produce the best result. Adding in the other features actually hurts the performance, but adding in the network features hurts the least, even though they overfit terribly.

Interestingly the overfitting with only the addition of the network features, but not with the addition of the non-network faculty suggests that the non-network features add a lot more nosie to the data, while the network features allow a very clear separation. Thus just adding the non-network features to the grant features yields the worst results.

The convergence parameters on the quasi-newton method that was being used in the linear classifier was played with extensively to try reduce the overfitting when using the grant and network features. Overfitting was reduced slightly down to 92%, but this did not change the testing accuracy much at all.

5.1.3 SVM

The SVM was a disappointment. It somehow kept breaking when the non-network faculty feature set was used, nor did it perform anywhere near as well as the linear classifier as it should. The regularization parameter was played with to no avail and thus only the random intermingling years test was completed which is shown in figure 4. This suggests a broken library[7] as the kernel used was a linear one.

[1] sum of distances to each vertex

[2] measures the expected number of times a node is traversed by a random walk averaged over all pairs of nodes

[3] how many shortest paths go through me

[4] based on average distance to each vertex

[5] the fraction of time that a random walk will spend at that vertex over an infinite time horizon

[6] how dense is my network around me

[7] or user error

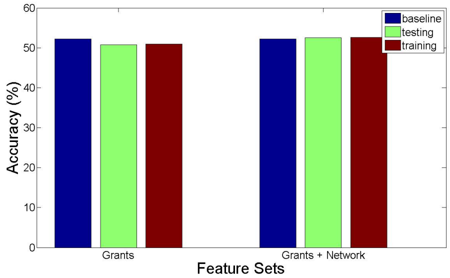

Figure 4: Results of testing on randomly intermingled years for the support vector machine.

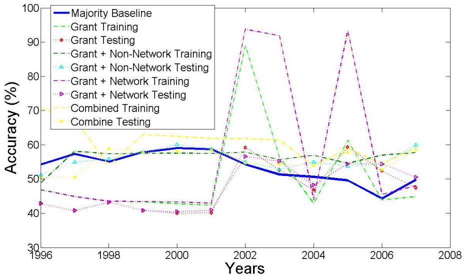

Figure 5: Testing on all years, training on the previous years for logistic regression

5.2 Training on previous years, testing on the next

5.2.1 Logistic Regression

Naturally this second set of tests should be expected to have a poor testing accuracy at the beginning as the training set size is rather small. This is clearly exhibited in figure 2, as all the testing accuracy lines start below the baseline and only start to creep over towards the end third of the years. The spikes in training accuracy seems to be related to the grant feature set. This unfortunately was not investigated further due to time constraints.

5.2.2 Linear Classifier

In figure 6 we see that in general all the features sets are above the baseline after the first half of the training data. Again the best performance came from just using the grant feature set, which infact was above baseline for all the years tested. It also exhibited a text book curve in terms of increased training data reducing the overfitting and increasing testing accuracy.

It is cool to see that towards the end all the classifiers had increasing accuracy, and were roughly achieving the same result, being a nice amount above baseline. This also indicates that the non-network features just required more training data to clear away any noise that was present in the intermingled year testing.

5.3 Top Features

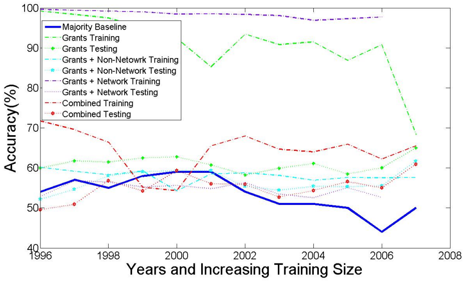

Figure 6: Testing on all years, training on the previous years for the linear classifier

Overall the top features from any of the sets turned out to be the bag of words on the title and sponsoring organization.

Table 1 shows the comparison between logsitic regression and linear classifier for the grant, and grant + network feature sets. Notice that logisistic regression was not able to discern betters weights for the features as the linear classifier. They are largely similar, but with the addition of the network features, the linear classifier is then able to find better weights for these largely similar features.

Overall the addition of the network features changes the top features to be more title oriented. No network features appear in the top twenty weighted features in either logistic regression or the linear classifier when using the grant + network feature sets.

Looking at table 2 the top features when using the non-network features sees the top being largely related to departments and the average minimum average amount award for a faculty member. The title or sponsoring organization features from the grants are not to be found in the top twenty weights.

The combined feature sets actually sees most of the top features come from the features in the network feature set. Interestingly one of its top features was on the publication coauthor graph, their betweeness centrality score, which was how often this person was on a path to all other people in the network.

6. CONCLUSION & FUTURE WORK

The bag of word model on the grant title and sponsoring organization does surprisingly well overall. But by just looking at the logisitic regression results one could hypothesize that the network features can help, but then looking at the linear classifier the network features actually hurt performance. This is probably due the overfitting that is happening, not allowing the model to generalize itself as well as just using the baseline features. More data or removing features would be the next approach to see whether the overfitting can be reduced.

From the results of the year to year tests, the non-network features showed that they were not infact a hinderance as all the classifier approached the same performance.

Table 1: Top features given by the classifiers with corresponding weights

Rank	*LC Grants*	*LC Grant + Network features*
1.	GSF.sponsor.org:Affairs 0.42320376196406856	GSF.sponsor.org:Merck 1.332017948078512
2.	GSF.sponsor.org:American 0.4193289572745211	GSF.project.title:Agreement 1.246513339384414
3.	GSF.sponsor.org:Veterans 0.4148158949296257	GSF.project.title:Proteases 1.1682290874186396
4.	GSF.sponsor.org:Defense 0.3324431609698573	GSF.project.title:Material 1.10557329991572
Rank	*LR Grants*	*LR Grant + Network*
1.	GSF.proposed.total: 3.4019115047822216E-6	GSF.project.title:Agreement: 2.1819922788681505
2.	GSF.sponsor.org:Institutes: 2.084470711054364E-11	GSF.sponsor.org :Merck: 2.1438229249679956
3.	GSF.sponsor.org:Health: 1.5289357472387933E-11	GSF.project.title:Material: 1.6967485373680784
4.	GSF.project.title:Material: 1.4695567254611942E-11	GSF.project.title:NSF: 1.6583620244115467

Table 2: Top features given by the classifiers with corresponding weights

Rank	*LC Grant + Non-network*	*LC Combined*
1.	LG.GDPT.totalAppliedInYear 0.005985296645086139	GDPT.ratio.aA.aR 0.00818726612441789
2.	SQ.GDPT.totalAppliedInYear 0.004965024091320833	NET.grantCA.BarycenterScorerW 0.0028962810755648995
3.	GPF.totalAwarded 8.772522148671472E-4	SQ.NET.pubCoAuthor.BCentralityW 0.001914286988455054
4.	GPF.MIN.avg.awarded 8.53257547026869E-4	LG.NET.grantCA.BarycenterScorer 0.002842374890139031

The change in top features across the sets was surprising, especially the ones picked by the combined feature set, suggesting that network features do infact hold some promise for use in machine learning. However it's clear that the bag of word model on the grant title and sponsoring organization proved to be the best features, when the classifier was able to weight them appropriately, in this case.

6.1 Future Work

There is lots that could be done to build on top of this work. For brevity the top three on my agenda are:

1. investigating the SVM issues, as this classifier should give just as good performance as the linear classifier.
2. looking at the network feature set and investigating the overfitting by removing features as it seems that classifier accuracy could be easily improved if overfitting was addressed.
3. approach the grant approval prediciton by clustering the data, and investigating whether prediction could be improved by using this clustering approach.

7. REFERENCES

[1] R. Burt. *Structural holes: The social structure of competition*. Belknap Pr, 1995.

[2] R. Burt. Structural Holes and Good Ideas 1. *American journal of sociology*, 110(2):349–399, 2004.

[3] J. OŠMadadhain, D. Fisher, S. White, and Y. Boey. The jung (java universal network/graph) framework. *University of California, Irvine, California*, 2003.