Research Funding Exploration

https://github.com/hansenrl/cs249project/

CS249 Project, Spring '14

Ross Hansen[1]
Utkarsh Jaiswal[2]
Mark Montoya[3]

Project Overview

In this project we analyzed patterns and trends in research funding from two datasets. The first dataset, ~8700 grant applications from the University of Melbourne, was used to build a classifier to predict successful applications. The second dataset, over 300,000 successful NSF grants from 1983-2008, only contains awarded grants and thus cannot be used for classification. However, the second dataset contains a long history as well as abstracts for each grant, allowing for a rich exploratory analysis.

Using these datasets, we explored questions such as:

- What factors contribute to the success of a grant application? (ie, what can be used to predict successful applications?)
- What fields of research see the most success in attracting funding?
- How are research funds distributed in the United States? (geographically, and by institution) Are they more concentrated in certain regions or institution types?
- Is the NSF consistently on either the leading edge of scientific research, or does it follow popular trends?

Datasets Used

Australia Funding Prediction Dataset

The funding prediction dataset was obtained from a kaggle competition and is available at http://www.kaggle.com/c/unimelb/data. It contains over 8700 applications from the University of Melbourne from 2005 to 2008 in a CSV format. For each application, it contains information about the application itself (the sponsoring organization, date of application submission, the grant amount, etc.), the research work (academic field, socioeconomic classification), and about the

[1] hansenrl@cs.ucla.edu
[2] uj@cs.ucla.edu
[3] msmontoya@cs.ucla.edu

researchers (role, country of birth, number of publications, previously successful grant applications, etc., for each researcher). The data does not contain abstract data or project titles, but does contain the status of the application - whether it was approved or rejected.

NSF Dataset

We looked at two separate datasets for NSF funding data, although collectively we refer to them as the "NSF Dataset". Our first dataset (obtained from research.gov[4]) consists of funding data from 2007-2013 that includes award information such as start and end dates, award amounts, funding agencies, PI and institution details, application abstracts and the primary Congressional district associated with the awardee. In addition, the dataset includes fields for project outcomes, including any publications both as a result of the award and as conference proceedings.

Our second dataset (obtained from nsf.gov[5]) spans twenty five years (1983-2008) worth of funding data. While it is similar to the first dataset in that it contains details about award expiration dates, award amounts, PI and institution details, and application abstracts, it lacks fields for project outcomes. However, it contains information about the areas of research each application is concentrated toward. Additionally, the relatively longer history of the second dataset vis-a-vis the first opens up other research possibilities as explained in the following section.

Analysis Attempted

Because of the two different types of data, separate analysis was done on each dataset.

Australia Funding Prediction Dataset

The end goal of analysis on the Australian dataset was to find a set of predictors that determine the success of an application, and then to construct a classifier that was effective at predicting if an application would succeed or fail. In the process of identifying the predictive features and constructing the classifier, extensive exploratory analysis was done on the data to uncover patterns that could be exploited for prediction.

Much of the exploratory analysis was in the form of pivoting the data around different features to determine if a feature had an affinity for success or failure. For example, we explored if grant values, prior investigator grant success, and number of investigator publications correlated well with application success. We also used the original data to construct meta-features to see if

---

[4] NSF dataset 2007-2013 - http://www.research.gov/research-portal/appmanager/base/desktop?_nfpb=true&_eventName=viewQuickSearchFormEvent_so_rsr

[5] NSF dataset 1983-2008 - http://www.nsf.gov/awardsearch/download.jsp

new features could be constructed that correlated with success or failure. See the notebook for details about specific trends that we identified with this exploration.

To do classification, four different learning algorithms were used: logistic regression, SVM, AdaBoost, and Random Forest. In the end, the Random Forest classifier had the best results, with an area under the ROC curve of ~0.92.

NSF Dataset

Since the NSF dataset does not contain information about failed applicants, it is not possible to build a classifier for it (due to a lack of negative training data). As a result, our primary goal here was exploratory analysis to discover patterns in funding by state, institution and investigator. In addition, we also looked at application success rates based on the time of the year that they were processed. Finally, we discovered some interesting trends in NSF funding using Google N-grams in order to determine whether it leads or lags popular technologies (our primary focus here was fields related to Computer Science).

The results of this analysis can be found in comprehensive format in the notebook itself, but we will list a couple of interesting results here.

- Georgia Tech is the most “up and coming” institution (based on % of NSF funding the institution manages to obtain over an extended period of time)
- The number of NSF awards in a particular field is closely proportional to the popularity of that field based on literature (Google NGrams) about 50% of the time, based on the research fields we investigated. It should be noted, however, that we only explored primarily CS related fields.
- When looking at whether the NSF awards for a reasearch area “lead” or “lag” general popularity, we found that the rise in popularity of a field in NSF awards lags the literature popularity 90% of the time. (“lag” in the sense that the NSF awards a smaller proportion of funds than the field is popular in Google ngrams, so the NSF is “lagging” general popularity in that area)
- When we normalized NSF popularity not only by number of awards in a year, but by the length of the award, the results did not change. The NSF still lags the vast majority of the time.

Experiences, Insights, and Lessons Learned

Although our team members had some experience in machine learning in the past, we had never worked with large “messy” datasets with many heterogenous features. We gained an appreciation for the amount of work required to import data into a usable representation and then clean and format the data into useful features.

A significant limitation of the 25-year NSF dataset was that it contained one XML file per award, as described in their schema[6]. With certain XML elements being repeated several times and no way of knowing their maximum number of occurrences over the 300,000+ files that we were working with, converting the data into pandas dataframes was more challenging than we initially expected.

Appending each XML file as a series to the dataframes involved a lot of copying, making our script memory intensive. As a result, we were forced to create a fixed length dataframe for each year's awards that employed padding/truncation as needed on the repeated elements of the XML files, and append data via lists instead. In addition, writing the dataframes to disk required experimental pandas methods (msgpack) in order to achieve a workable level of compression. The 25-year NSF dataset was also too big to fit entirely in memory on our desktop with 16GB of RAM, so we learned how to effectively work with that data in chunks in ipython.

Another important lesson that we learned was the importance of speed in the analysis, and practice using the built-in Panda and NumPy utilities such as groupby, apply, and ufuncs. Although it takes extra effort up-front to develop an elegant way of representing a desired analysis in the efficient utilities (for example, using a groupby, reduce operator, and NumPy ufunc), it results in a much faster and cleaner analysis than a hodgepodge of loops or verbose operations.

Finally, we have learned to be constantly skeptical of results obtained during exploratory analysis. During our analysis, we frequently got results, only to realize after thinking about it that the results are inconclusive (wasn't normalized properly, or we were forgetting to consider some confounding factors). This happened so frequently that we grew extremely cautious of any preliminary results. This caution is prudent in any type of exploratory analysis where interesting patterns and correlations can mislead researchers into mistaken conclusions.

Who Did What

Mark

- *Data* - Import of NSF data into ipython environment
- *Analysis* - Lead construction of NSF analysis notebook. Mainly exploratory analysis: distribution fitting, linear regression, developing a method for gauging the closeness of NSF funding to the popularity of a field in literature.
- *Documentation* - Worked on this summary, wrote documentation in NSF notebook.

[6] NSF's XML schema (25 year dataset) - http://www.nsf.gov/awardsearch/resources/Award.xsd

Utkarsh

- *Data* - Primarily charged with converting ~300,000+ XML files (one per NSF award for 1983-2008) into pandas dataframes. As detailed earlier, this was a non-trivial effort that entailed a lot of trial and error to get right given our memory constraints. In addition, a full run of the script across all years took about 5-6 hours to generate the final dataframes.
- *Documentation* - Worked on this summary and miscellaneous project documentation (GitHub)

Ross

- *Data* - Imported Australian Kaggle dataset into ipython environment and associated data wrangling
- *Analysis* - Lead construction of Australian Kaggle notebook. Exploratory analysis and trend finding as well as classification. Contributed analysis ideas and methods for NSF analysis.
- *Documentation* - Worked on this summary, wrote documentation in Australian Kaggle notebook.