Analysis of Political Contribution Data

Patrick Liu
University of California, San Diego
ppliu@eng.ucsd.edu

1. Introduction

Political parties within democratic systems have long sought ways to understand the demographics of their support. Historically, direct polls have been utilized to garner information on voter demographic along with issue-specific opinions, and are conducted before and throughout a campaign, and even after elections. The results of such polls shape political strategies.

With more data becoming publicly available, strategical data-mining has become en vogue. Most notably, public figures with strong analytical sense such as Dan Silver and Dan Wagner have been able to use data to make accurate election and voter predictions, with the former gaining notice for accuracy in predicting the state votes of the 2008 presidential election, and the latter being brought on as Chief Data Analyst for the Obama 2012 presidential election campaign.

As campaign finance relies a large portion on contributions [7], understanding the demographics of contributors may be useful as well. Political candidates in the United States are required to disclose contribution information, and the data is made public by the Federal Election Commission (FEC) [10]. The information includes the contributors name, information about their place of residence and employment, and date and amount contributed.

Here we attempt to use the information provided by the FEC to predict the political party for which the contributor will support. We hope to extract the useful features of such a predictor in order to gain insight about the demographics of these contributors.

2. Prior Work

Data mining to understand political demographics is not a novel idea, and has become an integral part of political campaigns [8]. Historically, this type of analysis is performed on poll information and ranges in goals from predicting the amount of support an individual will give a campaign to predicting the outcomes of the elections themselves.

With the volatile nature of politics and public opinion, and recent shifts in sensitivity towards privacy issues, much of this methodology is understandably not made public. One notable exception is the previously mentioned Nate Silver, who famously predicted the results of the 2008 presidential election with high accuracy. His efforts are focused heavily on optimization and curating the correct data to make accurate predictions, on which he runs linear regression analysis [9].

Data mining political data for research has also been done, though generally the goal in this field is for general understanding, as voting habits make interesting data sets for network analysis [1].

The particular data set examined in this paper is relatively obscure, and no relevant works directly relating to it could be found.

3. Data Preparation and Overview

I. Data Retrieval and Cleaning

The data used was retrieved from the FEC disclosure data portal [10] and contains the in-

formation about contributions during the 2012 presidential election campaign. A subset of that information was collected and are as follows:

- Candidate Name
- Contributor State
- Contributor Occupation
- Contribution Amount

Historically in the United States, individual states have been associated with political parties – enough so that states are often times referred as 'red' or 'blue' states depending on their political lean and 'swing' states (states that don't have a strong association with either party) are often strategically targeted by political campaigns to win their favor. Presidential campaigns also seek support from specific types of groups, oftentimes those who have views aligned with the candidates, through things like political promises (e.g. promises to focus on pollution regulation to gain support from environmental activists, or promises of tax cuts to gain the support of lower income groups). Our hope is that in choosing the features of state, occupation and contribution amount, we can capture the importance of these elements in determining which candidate an individual may support.

The self-reported nature of this data necessitated a way to clean the data. A large number of entries contained spelling errors in the state and occupation fields. More challengingly, occupation titles are largely unstandardized, and due to character length limitations, many entries contained abbreviations or shorthand (e.g. 'MD' or 'MGMT') and truncated words (e.g. 'FIREFIGH'). Entries with misspelled state codes were simply ignored. It was not immediately obvious which states were intended and these entries made up a small percentage of the overall data.

An initial attempt to categorize occupations into larger groups representative of their sector of work proved to be difficult. A rule-based approach to separate out the titles by keyword (e.g. if 'MANAGER' is in the title, separate out to a managerial occupation category) would oftentimes mis-categorize certain occupations (e.g. 'MEDIATOR' being categorized into a media occupation category).

Instead, a bag-of-words approach was chosen. The occupations were broken up by word and instances of the word counted, based on a corpus of overall words found in the entire dataset. Each occupation was then represented as a vector of word counts. As a way to reduce noise, words in the corpus that had instances less than 2 were discarded, as upon inspection these words were typically unusual spellings of more common words, rather than unique features of occupation titles.

The contribution amount was divided into two categories: small and large, for contributions under and above $200, respectively.

The candidate names were matched to their affiliated political parties, which were then used as labels for classification. Independent political parties were not well represented in the data and therefore not considered for classification.

After cleaning, the data was divided into a training, validation and test sets of sizes $88,825$, $6,263$ and $8,928$, respectively. In the training data, there are $69,119$ instances of Democratic contributors and $19,706$ Republican contributors, which make up 78% and 22% of the data, respectively. This imbalance in training examples will be discussed further in this paper.

II. Preliminary Examination of Data

We perform an exploratory analysis of the data to determine the viability of the proposed classifier. First, we examine the state origins of the contributors:

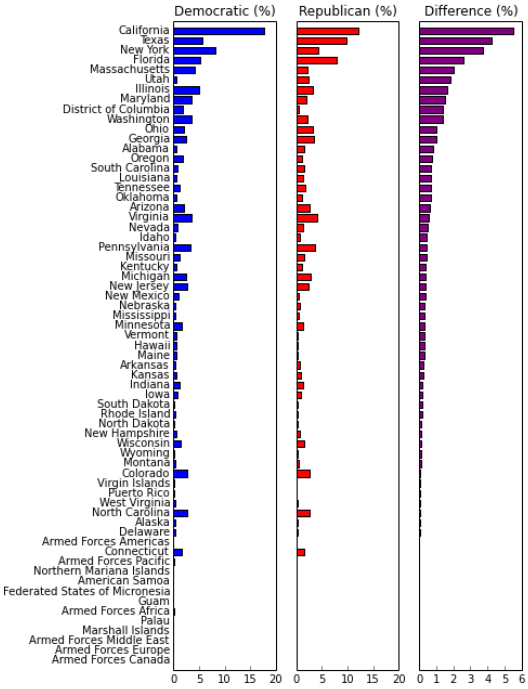

Figure 1: *Distribution of Contributors by State*

What we find is encouraging: when sorted by percentage difference, the top few states that have the largest differences generally favor the political party that won that particular state in the 2012 election (e.g. California and New York going to the Democrats and Texas going to the Republicans). An interesting exception is Florida, which was known to be a swing state. Following our previous logic, we would expect Florida to go to the Republicans, but in the election was won by the Democrats.

Next, we move on to looking at the differences in contribution sizes between the two parties:

Table 1: *Contribution Size by Party*

Contribution Size		
<$200	>$200	Party
60,436	8,683	Democratic
11,768	7,938	Republican

Again, the results are encouraging. There seems to be a discernible difference in the size of contributions with respect to different political parties. The Democrats seem to have a larger percentage of contributions over $200.

Lastly, we examine the word counts from our previously defined bag-of-words. The most common words found in the occupation field that were unique to Democratic contributors were these:

1. EDUCATOR
2. WORKER
3. SOCIAL
4. PSYCHOLOGIST
5. SCIENTIST
6. EDITOR
7. RESEARCH
8. PROGRAM
9. SCHOOL
10. DISABLED
11. PRODUCER
12. COORDINATOR
13. CLINICAL
14. COUNSELOR
15. LIBRARIAN
16. INSTRUCTOR
17. MUSICIAN
18. EDUCATION
19. PSYCHOTHERAPIST
20. ADMIN

For the Republican contributors, the most common unique words were these:

1. INSURANCE
2. FARMER
3. CONTRACTOR
4. INVESTMENT
5. INVESTOR
6. BROKER
7. SMALL
8. CHAIRMAN
9. GENERAL
10. PILOT
11. CONSTRUCTION

12. PARTNER
13. DRIVER
14. MECHANIC
15. COMMERCIAL
16. CLERK
17. PRIVATE
18. PROPERTY
19. REPRESENTATIVE
20. MGR

While tempting, it may be better advised to not draw many conclusions from these lists. To reiterate, the occupation field in the data was the one with the greatest variability and non-conformity to any standard. Along with that, breaking an occupation title into individual words may remove some of the context.

However, many of these words carry information irregardless of context (e.g. FARMER or MUSICIAN). It can also be said that at a very shallow overview, the lists do seem to represent distinct sectors of careers, with the Democratic list on the side of education, science and the arts, and the Republican list on the side of business, and trade and labor type careers.

We make the claim that these lists further support the notion that this data can be used to separate between the two parties, though in the case of occupation data, interpretation is not trivial.

4. Methods

I. Evaluation

We evaluate our model based on its accuracy of prediction. As a baseline, we will compare the results of our classification with a classifier that assigns labels stochastically based on the class probabilities, or the fraction of labels in our training data.

We will then evaluate the viability of the features deemed significant by our classifier by comparing them to our initial hunches based on the preliminary exploration of the data, and also to what is known or established through other means, such the election results. For the state features, this may be done by simply looking at the popular vote by state for the 2012 presidential election. For the contribution amount and occupation features this may require some creative interpretation, which we will try to relate to the parties campaign platforms.

II. Classification

The classifier being used is Multinomial Naive Bayes. This particular classifier was chosen for its effectiveness in high-dimensional data, and also the ability to interpret the parameters of the fitted model. The use of Multinomial Naive Bayes in the classification of textual data is well studied [4], and its generally good performance in this field motivated the decision to use this model. The decision was further motivated by the relatively light computational cost in comparison with other methods, allowing for scalability.

Naive Bayes is known to operate on the assumption of feature independence [4]. For our problem, the reality is that there is indeed some real world correlation between a persons state of residence, occupation and income level. However, modeling this information with the relatively small sub-sample that is this data in comparison with the overall population of the United States would be difficult. The argument can also be made for interdependence of words in a persons occupation title (e.g. 'REAL' being correlated with 'ESTATE'). Again, modeling this dependence with the relatively small corpus size and high variability in occupation titles would be a challenge. For future work, this is an area that could possibly improve the performance of our classifier, but for this paper we continue under these assumptions.

The Naive Bayes model has relatively few pa-

rameters to tune, namely a Laplace/Lidstone smoothing parameter α and the choice to fit a prior distribution based on the class probabilities. Due to the skewed nature of our training data we will choose to fit this prior distribution, but for the sake of comparison we show the results for abstaining from this option as well. The smoothing parameter α will be chosen based on classifier performance on the validation set.

A common optimization in text-mining is to use the term frequency–inverse document frequency metric rather than raw counts [4] to adjust for frequency of word appearance. Our corpus is composed of 'documents' of mostly one to three word phrases where words are highly unlikely to be repeated, and so we continue without this optimization.

Other models considered for classification were Linear SVM and k-Nearest Neighbors. These models did not perform as well as the Naive Bayes model in accuracy. The Linear SVM model performed similarly to Naive Bayes, but ultimately not as well. Another motivation for choosing the Naive Bayes model over Linear SVM is the interpretability of the fitted models parameters, which will be discussed in the section Feature Extraction.

The k-NN model does not intrinsically have any interpretable parameters. Methods returned by a search for feature selection with k-NN involve constructing multiple models with subsets of the original features [11]. In combination with the longer run-time of k-NN due to the nature of the method, this was deemed to be difficult to scale. Additionally, initial attempts to use k-NN yielded low classification accuracy. This could be explained by the sparsity of the feature vectors and the practically completely categorical nature of the data (occupation titles are unlikely to repeat words), where traditional distance measures may under-perform.

III. Feature Selection

Beyond classification, our goal is to understand how a contributors resident state, occupational type, and income level contribute to their political affiliation. In the context of our classifier, we seek to determine the importance of these individual features in the resulting decision function.

In our Multinomial Naive Bayes model, we achieve this by examining the calculated likelihoods of individual features x_i given a class C or $P(x_i|C)$. The individual likelihoods are unlikely to have useful information on their own, so to interpret these likelihoods we compare them between the two classes on a feature by feature basis.

To further discuss the decision to forgo using Linear SVM, we examine methods of feature extraction for this model. In the case of Linear SVM, the fitted parameter is a vector representing the separating hyper-plane. The relative absolute size of each dimensions coefficient in this vector can be argued to be indicative of that features importance. This idea is applied to the Recursive Feature Elimination (RFE) method for feature selection, which ranks subsets of features based on the impact to classification performance when removed. The combination of Linear SVM and RFE is well studied and most notably used in solving problems of gene importance in disease studies [2][3].

For this problem, it was found that using the RFE method in combination with the Linear SVM model gave results that were a mix of intuitive given the information gathered in the initial data exploration, and noisy in the sense that under-represented features in the training data were also ranked highly.

5. Results

I. Classification

In tuning the Laplace/Lidstone smoothing parameter α, we perform an uninformed search over several potential values, reporting the accuracy on the validation set:

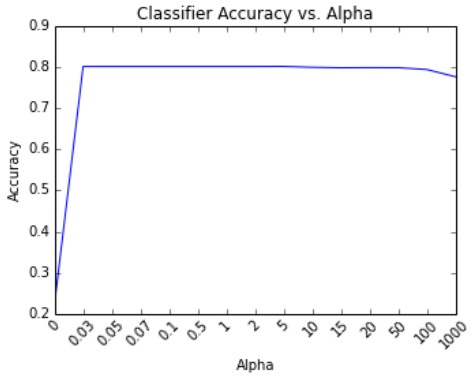

Figure 2: *Tuning the Smoothing Parameter α*

It was found that an α value between 0.03 and 5 gave the best performance, with only slight decline as the numbers increase. A smoothing parameter of 0, or no smoothing at all, gave the worst performance with an accuracy of 22%. We fix α to be 1, as in accepted practice [12].

As mentioned earlier, the imbalance in training data motivates the fitting of a prior distribution based on the class probabilities to our model. Without such action, the classifier does not perform as well, giving an overall accuracy of 68% in comparison to 80% with on the validation set.

On the test set, our model was found to have an accuracy of 81%. To compare with the aforementioned baseline of random prediction, $10,000$ sets of predictions were generated by sampling from a random discrete variable with probabilities equal to the class probabilities, or the fraction of the overall training data for either label. This method was found to have an accuracy of 65%, with a standard deviation of 0.4%.

To further inspect the performance of the Naive Bayes model, we create a confusion matrix:

Predicted \ Actual	R	D
R	652	385
D	1,281	6,610

The per-class accuracy is 63% for the label 'Republican' and 84% for the label 'Democrat'.

II. Feature Selection

In examining the features when ranked as described in the Methods section, we find some interesting results that support the ideas gathered from the data exploration. After filtering for features with a log-likelihood less than -9 in both classes (under-represented features in both classes), we are left with 274 features to compare. We report the top 20 features with the largest disparity in likelihood, first favoring the Democratic Party:

1. SOCIAL
2. EDUCATOR
3. PSYCHOLOGIST
4. PSYCHOTHERAPIST
5. LIBRARIAN
6. PROFESSOR
7. CLINICAL
8. WORKER
9. SCHOOL
10. PRODUCER
11. RESEARCHER
12. WRITER
13. EDUCATION
14. COUNSELOR
15. MUSICIAN
16. SCIENTIST
17. FACULTY
18. EDITOR
19. COMMUNICATIONS
20. DC (state code)

And then favoring the Republican Party:

1. CHAIRMAN
2. INVESTMENT
3. FARMER
4. UT (state code)
5. INVESTOR
6. SELF
7. PRIVATE
8. MECHANIC
9. INSURANCE
10. INFORMATION
11. size of contribution (>$200)
12. CONTRACTOR
13. CONSTRUCTION
14. COMMERCIAL
15. REAL
16. ESTATE
17. ID (state code)
18. GOVERNMENT
19. ND (state code)
20. PRESIDENT

A majority of the features shown in these two lists are features that were projected to be useful in our initial data exploration. Considering the nature of the Naive Bayes model, this result is not surprising, but it is reassuring.

To more closely examine the ranking of the state features in particular, we again report the states with the largest disparity in log-likelihood, along with the result of that state in the 2012 election. The top 10 states favoring the Democratic Party as found by our method of feature extraction are:

1. Washington, D.C. (D)
2. Vermont (D)
3. Maine (D)
4. Hawaii (D)
5. Massachusetts (D)
6. New York (D)
7. Maryland (D)
8. Oregon (D)
9. Rhode Island (D)
10. New Mexico (D)

The top 10 favoring the Republican Party:

1. Utah (R)
2. Idaho (R)
3. North Dakota (R)
4. Alabama (R)
5. South Dakota (R)
6. Louisiana (R)
7. Oklahoma (R)
8. Mississippi (R)
9. Nebraska (R)
10. South Carolina (R)

Again, the high correspondence of states ranked by our feature selection and the eventual outcome in the 2012 election is what was expected, and the results are encouraging that some useful features are being chosen.

To convince ourselves of the utility of occupation title words chosen by feature selection, we look to the political platforms for each party. In examining the lists, we again draw similar conclusions as from the initial data exploration: the Democratic contributors appear to be in occupations related to education, science and research, and personal care. In 2009, President Barack Obama signed into law the Recovery Act. Among other things, one of the states objectives of this act was to invest more heavily in education and health [6]. The Republican contributors on the other hand appear to be in occupations more related to business, management, and skilled labor or tradesmen. The Republican Party Platform mentions supporting business and entrepreneurship, and mentions agriculture and farmers directly [5].

6. Conclusion

In this study, we examine data provided by the Federal Election Commission detailing political contributions. From the data, we selected the features of the contributors state of residency and occupation, along with the contribution

amount. From this, we hoped to predict the political party that the contributor supports.

We find that a Multinomial Naives Bayes classifier is capable of predicting this to an acceptable degree of accuracy. Additionally, we find that the parameters of the fitted Naive Bayes model allow us to extract useful features, which we can explain to a certain extent by examining each states pattern of voting and the platform of each political party.

References

[1] A. Jakulin, W. Buntine, T. M. La Pira and H. Brasher.
Analyzing the U.S. Senate in 2003: Similarities, Clusters, and Blocs
Political Analysis, 2009, 17 (3): 291-310.

[2] Ying Yu.
SVM-RFE Algorithm for Gene Feature Selection

[3] I. Guyon, J. Weston, S. Barnhill, M.D. and V. Vapnik.
Gene Selection for Cancer Classification using Support Vector Machines

[4] J. Rennie, L. Shih, J. Teevan, and D. R. Karger.
Tackling the Poor Assumptions of Naive Bayes Text Classifiers
ICML, 2003

[5] We Believe in America: 2012 REPUBLICAN PLATFORM
`https://www.gop.com/platform/`

[6] The American Recovery and Reinvestment Act
`http://www.recovery.gov/arra/About/Pages/The_Act.aspx`

[7] J. D. Salant.
Romney Shunning Federal Funds in Post-Watergate Election
Bloomberg, 2012

[8] S. Issenberg.
How President Obamas campaign used big data to rally individual voters.
MIT Technology Review, 2012

[9] N. Silver.
Methodology
`http://fivethirtyeight.blogs.nytimes.com/methodology/`

[10] F.E.C.
2012 Presidential Campaign Finance
`http://www.fec.gov/disclosurep/PDownload.do`

[11] S. Li, E. J. Harner, and D. A. Adjeroh.
Random KNN feature selection - a fast and stable alternative to Random Forests
BMC Bioinformatics, 2011, 12:450.
`http://www.biomedcentral.com/1471-2105/12/450`

[12] D. Jurafsky and J. H. Martin.
Speech and Language Processing (2nd ed.)
Prentice Hall, 2008 , ISBN 978-0-13-187321-6.