Vote Prediction in the House of Representatives

Benjamin Shulman, Jisha Kambo, John Oliver

December 13, 2012

1 Introduction

The federal government of the United States is organized into three branches: Executive, Legislative and Judicial. The legislative branch (Congress) is responsible for sponsoring bills and voting on whether bills should become law. The United States Congress is divided into two sections, the House of Representatives and the Senate. The House of Representatives is made up of 435 representatives, each of which are elected every two years. The Senate is made up of 100 senators, 2 from each state and are elected on staggered six year cycles. The general process for a bill becoming a law is as follows. Each bill has a sponsor and is introduced to the House of Representatives, the bill then goes to committee and then is revised, researched and then sent to the House. The bill is debated by the representatives, and changes and amendments are recommended. The bill is then voted on by the members of the House. If it passes it is sent to the Senate where it goes through a similar process. If it is passed by the Senate it is sent to the president to be signed.

A new law passed by Congress may have vast effects on American citizens and corporations, affecting revenue, assets, income, rights and freedoms. If citizens and corporations could know whether a bill would be passed by Congress before it was voted on it would give them advance warning such that they can lobby for changes to the bill, lobby for it to be shut down or passed or prepare themselves for the resulting effects. Further, if an individual representative or senator's vote could be predicted for a given bill it could give insights into the actual beliefs and platform of that congressperson. With these reasons in mind we will tackle the problem of vote prediction in the House of Representatives. We will approach the problem by treating each representative as a separate prediction problem and then use those predictions to predict whether a bill will pass the House of Representatives. There has been very little exploration of this space, the only related work being done using the text bills to predict roll call votes by extending the ideal point model (Gerrish and Blei). Our work will focus not on the bill text, but on metadata about the bill and does not extend the ideal point model.

In the following sections we will discuss our methodology for modeling and vote prediction. We examine and discuss our data extraction and feature selection. Further we show that bill metadata can be used to accurately predict the votes of individual representatives in the House using Support Vector Machines and

compare our results against a baseline model for how representative's vote.

2 The Problem

We are specifically interested in two questions:

1. Can we predict an individual member of the House of Representative's vote on a bill with high accuracy in comparison to a baseline prediction model?

2. Can we use prediction of individual representative's vote to successfully predict whether a bill will be passed by the House of Representatives?

Other questions we were interested in, but do not directly address in this publication are:

1. What features of a bill are most important to members of the House when voting?

2. What machine learning algorithms are most effective for this problem?

3 Algorithms and Methods

To address our first question we focused primarily on using support vector machines to predict how each Representative would vote. For each representative we trained, tuned using a validation set and tested a separate support vector machine. To compare the success of support vector machines in predicting representatives' votes we used a baseline hypothesis which predicted that if a representative was of the same party as the sponsor of the bill he or she would vote "Yes" or if not, would vote "No." To predict whether a bill would be passed or not we did not use a standard machine learning algorithm as we had already done the prediction part using the individual support vector machines that predicted representatives' votes. Instead we used a simple algorithm that took a sum of predicted votes, where each vote is weighted by the accuracy of the representative's support vector machine on the test set for that representative.

4 Methodology

4.1 Data

We took data from www.govtrack.us, a website promoting government transparency. From Govtrack's database we pulled data about all current representatives. We excluded five representatives from our experiment as there were problems with their information in the database, leaving us with 430 representatives. We used Govtrack's database to extract the voting records of these 430 representatives. For each representative we pulled all bills which were being voted on for passage which they had ever voted on. Each bill represented an example for a representative. The data from the database about each bill included information about the status of the bill; when it was introduced; congress number; when it was voted on; its title; and information about the sponsor of the bill such as name, district, and party.

4.2 Feature Selection

Feature set:

- Party of bill sponsor
- Name of bill sponsor

- District of bill sponsor
- Gender of bill sponsor
- Date bill sponsor joined House of Representatives
- Whether bill sponsor has Twitter
- Whether bill sponsor has a website
- Whether sponsor has a nickname
- Length of time bill has been alive
- Congressional Session
- Year the bill was introduced
- Day the bill was voted on
- Month the bill was voted on
- Year the bill was voted on
- Year the bill was voted on mod 2
- Year the bill was voted on mod 4
- Year the bill was voted on mod 6

From the information we pulled from each bill we selected a set of features we wanted to consider in our models. The first was the party of the bill sponsor, we believe that representative's are likely to be have party bias and thus this is an important feature. We also used the sponsor's gender, as some representatives may have gender bias. Further we chose the sponsor's district and name as well, as representative's vote may be dependent upon certain districts or sponsors. For each representative, a bill sponsor's district and name were each a binary feature. Meaning if there were ten training bills for a representative, there may be up to ten features for sponsor name and up to ten for sponsor district. Final sponsor information we used as features were the date the sponsor joined the House, whether he or she had a Twitter, whether he or she had a website, and finally whether he or she had a nickname–each of these features were considered important as they could be indicators of the influence, popularity and experience of the sponsor which may be relevant to the voting representative.

We included several other pieces of information about the bill. We extracted seven features from the date of the vote on the bill: the day of the vote, the month of the vote, the year of the vote, the year mod two, the year mod four and the year mod six. The day, month and year may all matter as priorities change given the time of year. The three modded year features align with voting cycles: representatives are elected every two years, president every four and senators every six. Thus if a representative was running in one of those elections, the time in relation to the cycle may affect his or her vote. Finally we included the congressional session in which the bill was voted on and the length of time the bill was active in days before it was voted on. These two features were included as the length of time a bill is debated for may effect votes as well as the session in which it is voted on. All of these features became the feature set we used in our models.

4.3 Model Selection

The machine learning model we chose to use for representative prediction tasks was a support vector machine. We treated each representative as a separate prediction task. Thus

we had a separate data set for each member of the House. We split each data set into training, validation and testing sets. Thirty percent of the data for each member was used for testing, the remaining seventy percent was split between training and validation, with eighty percent going to training and twenty percent to validation. Then for each representative we trained, and tuned a separate support vector machine, using the validation set to pick the best C value from five options $[0.0001, 0.001, 0.01, 0.05, 0.1]$. Thus we had 430 separate support vector machines.

4.4 Statistical Analysis

5 Discussion

6 Related Work

7 Future Work

8 Conclusion