Using Active Learning to Label Data Efficiently

Joe Patten

Abstract

Labelling data can be expensive or even intractable. This paper seeks to explore some active learning approaches in order to efficiently use an expert to better predict unlabelled data. I use Naive Bayes in order to classify law cases that have been judged or settled. I use random sampling, least confident sampling, and margin sampling to show how each performs on this dataset. I would like to acknowledge my professor, Jana Doppa, and TA, Aryan Deshwal, for all they taught this semester. I would also like to acknowledge my advisor, Ben Cowan, for working with me on whistleblower research.

Introduction

The ultimate objective of this paper is to explore active learning methods in order to classify a large dataset. The dataset in question contains about 15,000 law cases that are in need of labels. Each case contains a summary of a few paragraphs. We need labels for this dataset as we want to focus on certain "classes" of cases. It would be inefficient and expensive (in terms of man hours) for someone to sit down and classify each and every case. Thus, it would be wise to explore algorithms and techniques that would allow us to accurately classify these cases without having to label a ton of cases by hand. Active learning helps us to do just that.

Qui-Tam Cases

Currently, we are trying to estimate the effects of corporate whistleblowers on government savings. In the last few years, there has been an increase in studies done on whistleblowing; however, the majority of these papers have a law focus (Engstrom, 2012, 2014), with very few actually analyzing the economic effects. We hope to fill in this gap in the literature.

We are still waiting on a firm in D.C. for the dataset mentioned in the introduction. Although we don't have access to this data, we can make a similar dataset by scraping reports issued by the DOJ in order to test our model.

Model

Formally, I want a model that can classify each case with one of the 3 labels: tax, health care, and general fraud. There are some imperfections with this approach. For one thing, these three labels are fairly general. For example, the general fraud label can be further broken down into housing, financial, defense, and corruption. Also, there are some cases that could be classified with multiple labels. I will first start by evaluating my model by using all of my training data. Then, I will compare how random sampling and uncertainty sample perform. Naive Bayes will be used to predict class labels. Although there are "fancier" algorithms out there, I will use Naive Bayes as it still performed decently on my data when compared with these other algorithms [1]. Also, the main focus of this paper is to explore some active learning methods to classify data. Another nicety is that active learning can be used for various algorithms and models.

Experiments and Results

My ultimate objective is to come up with a model that is able to accurately predict the class of a case, with the least number of hand-coded labels. I will start on a much smaller dataset that I scraped from the Department of Justice website. The dataset contains just over 500 settlements and judgements. Each case contains a summary paragraph that explains the case. There are three classes in the dataset: tax, health care, and general fraud. I have gone through and hand-labelled each case. Figure 1 shows the number of cases by type.

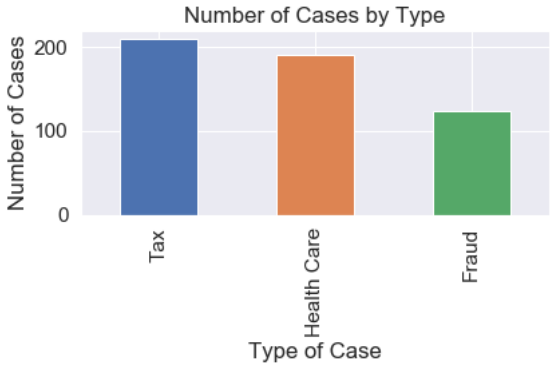

Figure 1: Cases by Type

The words of the case summaries will make up the features of the model. That means that first I need to vectorize the summary of each case summary. The following excerpt is an example of a case summary:

```
Fourteen hospitals located in New York, Mississippi, North Carolina, Washington,
```

[1] Boosting, Random Forests, and Bagging gave similar results as the Naive Bayes Classifier

```
Indiana, Missouri and Florida have agreed to pay the United States a total of
more than $12 million to settle allegations that the health care facilities
submitted false claims to Medicare, the Justice Department announced today.
```

I would classify the above summary as being the Health Care type. I will use Naive Bayes to classify each case. I will compare random sampling with uncertainty sampling. As all of the data is available, I will be using pool-based sampling. In both cases, I have split the data into training and testing using an 80:20 split. Before exploring active learning, I will use all of the training data to fit my model. This leads to a test accuracy of 94.29%. I believe that using accuracy is a fairly good evaluation metric as there are only 3 classes, and each class is well represented. However, the confidence matrix in figure 2 gives us a better look of how the model performs. The appendix also includes precision and recall scores for each of the class labels. Moving forward, I will use accuracy to evaluate my models.

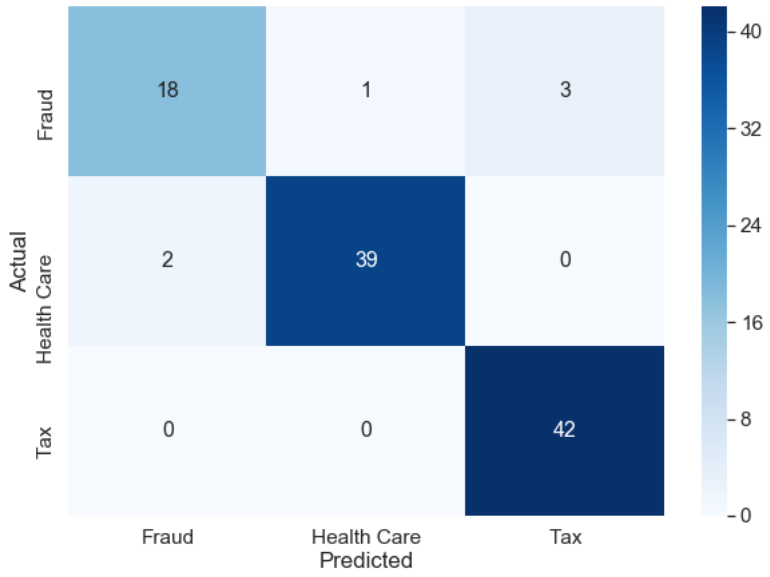

Figure 2: Confidence Matrix on test set given model is trained on the whole training set

Random Sampling

First I start with 30 random examples in the training data, and then add examples by pulling from the training set randomly. The model is then trained on the updated dataset. Figure 3 shows the accuracy when training examples are randomly sampled.

Uncertainty Sampling

Again, I start out with 30 random examples from the training set. I will follow Settles (2009), and will use uncertainty sampling by picking the example from the unpicked training set whose prediction is

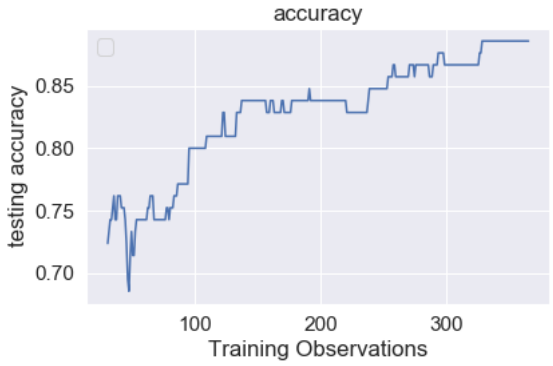

Figure 3: Accuracy given random sampling

the least confident and adding it to the updated dataset:

$$x^*_{LC} = argmax_x 1 - P_\theta(\hat{y}|x)$$

where $\hat{y}$ is the label with the highest probability given the model. Figure 4 shows the accuracy when training examples are selected given least confident sampling.

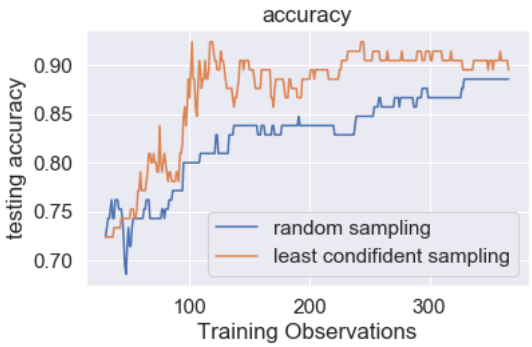

Figure 4: Accuracy given least confident sampling

This shows that using least confident sampling provides us with more accuracy than just random sampling. However, can we improve even more? It is often beneficial to not just use information on the most probable label, but also information on the second most probable label. Margin sampling leverages this information. Training examples are added to the model based on the argmin of the difference between the probabilities of the two highest predicted labels for each example:

$$x^*_M = argmax_x P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x)$$

Figure 5 shows the results from using margin sampling.

It appears that using margin sampling performs even bettering than using least confident sampling. This should definitely be a tool in every data scientist's toolkit.

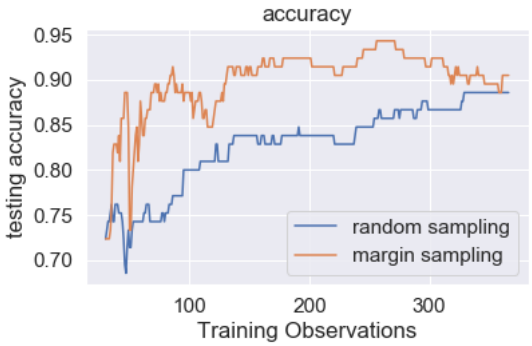

Figure 5: Accuracy given Margin Sampling

Conclusion and Further Work

Least confident sampling and margin sampling are just two active learning techniques that yielded improvements when compared with random sampling. Specifically, I showed that testing accuracy was larger for most number of observations when using either least confident or margin sampling (sometimes even twenty percentage points of accuracy increase). As data becomes more and more available, labelling unlabelled data becomes more and more expensive. Active learning is thus an efficient way of making the most out of unlabelled data. Once I have the large dataset of judgements and settlements, I will definitely be using active learning to classify the law cases in an effective manner. I will also look at different techniques (for example query by committee, expected model change, expected error reductionm,etc), and will also use other algorithms or models besides Naive Bayes.

References

Buraimo, B. (2008). Stadium attendance and television audience demand in english league football. *Managerial and Decision Economics*, 29(6):513–523.

Buraimo, B. and Simmons, R. (2009). A tale of two audiences: Spectators, television viewers and outcome uncertainty in spanish football. *Journal of economics and business*, 61(4):326–338.

Buraimo, B. and Simmons, R. (2015). Uncertainty of outcome or star quality? television audience demand for english premier league football. *International Journal of the Economics of Business*, 22(3):449–469.

Engstrom, D. F. (2012). Public regulation of private enforcement: Empirical analysis of doj oversight of qui tam litigation under the false claims act. *Nw. UL Rev.*, 107:1689.

Engstrom, D. F. (2014). Whither whistleblowing? bounty regimes, regulatory context, and the challenge of optimal design. *Theoretical Inquiries in Law*, 15(2):605–634.

Fort, R. and Lee, Y. H. (2006). Stationarity and major league baseball attendance analysis. *Journal of Sports Economics*, 7(4):408–415.

Mills, B. and Fort, R. (2014). League-level attendance and outcome uncertainty in us pro sports leagues. *Economic Inquiry*, 52(1):205–218.

Paul, R. J., Wachsman, Y., and Weinbach, A. P. (2011). The role of uncertainty of outcome and scoring in the determination of fan satisfaction in the nfl. *Journal of Sports Economics*, 12(2):213–221.

Paul, R. J. and Weinbach, A. P. (2007). The uncertainty of outcome and scoring effects on nielsen ratings for monday night football. *Journal of Economics and Business*, 59(3):199–211.

Rascher, D. A. and Solmes, J. (2007). Do fans want close contests? a test of the uncertainty of outcome hypothesis in the national basketball association. *A Test of the Uncertainty of Outcome Hypothesis in the National Basketball Association (June 15, 2007)*.

Rottenberg, S. (1956). The baseball players' labor market. *Journal of political economy*, 64(3):242–258.

Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Tainsky, S. and Winfree, J. A. (2010). Short-run demand and uncertainty of outcome in major league baseball. *Review of Industrial Organization*, 37(3):197–214.

Tainsky, S., Xu, J., and Zhou, Y. (2014). Qualifying the game uncertainty effect: A game-level analysis of nfl postseason broadcast ratings. *Journal of Sports Economics*, 15(3):219–236.

Appendix

	Case Type	Precision	Recall
0	Health Care	0.975000	0.951220
1	Tax	0.933333	1.000000
2	Fraud	0.900000	0.818182

Table 1: Precision and recall for Naive Bayes using the whole training set.

Code available upon request.

1 Introduction

There is a popular hypothesis (UOH) that a more balanced sporting competition leads to greater interest (Rottenberg, 1956). In other words, the UOH says that spectators generally prefer sporting events where two teams are of similar ability. Using Monte Carlo simulations to forecast a team's playoff probability in MLB, Tainsky and Winfree (2010) find that outcome uncertainty has no impact on match attendance. Mills and Fort (2014) and Paul et al. (2011) find evidence in favor of the UOH for NFL games. In the NBA, Rascher and Solmes (2007) and Mills and Fort (2014) find that more balanced matches increase stadium attendance. There have been several studies on multiple sports that have shown mixed results. Most of the literature has focused on stadium attendance, however, recent studies have started evaluating TV audiences. Multiple studies find the evenly matched games bring in larger TV audiences for the NFL, the English Football League, and the Spanish Football Primera League ((Buraimo, 2008), (Paul and Weinbach, 2007), (Tainsky et al., 2014), (Buraimo and Simmons, 2009)). This study will be the first of its kind that evaluates competitive balance and the superstar effect in esports. As esports is an area of rapid growth, I believe this study will shed some light on this unstudied area.

1.1 Uncertainty of Outcome

Uncertainty of Outcome (UO) can be measured in a few ways. Past studies have primarily used betting odds or predicted outcome (usually using a probit model) to calculate the probability each team in a match would win. Playoff Uncertainty (PU) has also been considered (Fort and Lee, 2006). I plan on constructing estimates for UO and PU by using Monte Carlo methods (which I have already played with for one season of an esport). Using these estimates, I will evaluate how these estimates impact viewership. Peak viewership for matches will be recorded from Twitch.tv. This is a platform where users stream games. Several esports leagues and tournaments are also streamed on this platform.

1.2 Superstars

In addition to uncertainty of outcome being a factor in viewership, I also believe that viewership is affected by the superstar effect. Buraimo and Simmons (2015) find that although outcome uncertainty had little effect on TV audiences for the Premier League, games in which superstars played on one of the teams boosted TV viewership. I will find the superstars for each esport, and evaluate the impact on online viewership. What is the definition of a superstar? Some say superstars are individuals get a large share of profits for the industry they work in. I will use player earnings as well as Twitch.tv logs to measure superstardom.

2 Data

Viewership data will be taken from Twitch.tv. There are a number of sites that store historical viewership data. I have contacted one site (sullygnome.com) that collects this data every 15 minutes. During a match on Twitch, viewers can comment and chat while an esports event is live. I will scrape these chatlogs in order to get a better picture of which players are superstars.

Player earnings data comes from esportssearnings.com. This site records the amount pro players receive from esports tournaments and league play. The highest earning player is currently N0tail, who has received nearly $7,000,000 in tournament and league play playing the game Dota 2.

3 Methods

I will evaluate the following function:

$$\begin{aligned}\text{Twitch Viewership}_{jt} = f(&\text{Outcome Uncertainty}_{jt}, \text{Superstar}_{jt}, \text{Home Team Performance}_{kt}, \\ &\text{Away Team Performance}_{lt}, \text{Match Characteristics}_{jt})\end{aligned}$$

where j denotes the match, t denotes the time variable, and k and l denote the home and away teams. Superstar is a dummy variable indicating whether there is a superstar in either team in the match.

References

Buraimo, B. (2008). Stadium attendance and television audience demand in english league football. *Managerial and Decision Economics*, 29(6):513–523.

Buraimo, B. and Simmons, R. (2009). A tale of two audiences: Spectators, television viewers and outcome uncertainty in spanish football. *Journal of economics and business*, 61(4):326–338.

Buraimo, B. and Simmons, R. (2015). Uncertainty of outcome or star quality? television audience demand for english premier league football. *International Journal of the Economics of Business*, 22(3):449–469.

Engstrom, D. F. (2012). Public regulation of private enforcement: Empirical analysis of doj oversight of qui tam litigation under the false claims act. *Nw. UL Rev.*, 107:1689.

Engstrom, D. F. (2014). Whither whistleblowing? bounty regimes, regulatory context, and the challenge of optimal design. *Theoretical Inquiries in Law*, 15(2):605–634.

Fort, R. and Lee, Y. H. (2006). Stationarity and major league baseball attendance analysis. *Journal of Sports Economics*, 7(4):408–415.

Mills, B. and Fort, R. (2014). League-level attendance and outcome uncertainty in us pro sports leagues. *Economic Inquiry*, 52(1):205–218.

Paul, R. J., Wachsman, Y., and Weinbach, A. P. (2011). The role of uncertainty of outcome and scoring in the determination of fan satisfaction in the nfl. *Journal of Sports Economics*, 12(2):213–221.

Paul, R. J. and Weinbach, A. P. (2007). The uncertainty of outcome and scoring effects on nielsen ratings for monday night football. *Journal of Economics and Business*, 59(3):199–211.

Rascher, D. A. and Solmes, J. (2007). Do fans want close contests? a test of the uncertainty of outcome hypothesis in the national basketball association. *A Test of the Uncertainty of Outcome Hypothesis in the National Basketball Association (June 15, 2007)*.

Rottenberg, S. (1956). The baseball players' labor market. *Journal of political economy*, 64(3):242–258.

Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Tainsky, S. and Winfree, J. A. (2010). Short-run demand and uncertainty of outcome in major league baseball. *Review of Industrial Organization*, 37(3):197–214.

Tainsky, S., Xu, J., and Zhou, Y. (2014). Qualifying the game uncertainty effect: A game-level analysis of nfl postseason broadcast ratings. *Journal of Sports Economics*, 15(3):219–236.