Auto-Calibration of Review Bias through Knowledge

Michael Vander Meiden
Robotics Institute
Carnegie Mellon University
mvanderm@andrew.cmu.edu

Abstract

The paper explores the problem of reviewer bias during the rating of submissions to academic conferences. We introduce and test a strategy to combat this bias. The strategy involves providing the reviewers with a previously collect distribution of ratings and the desired distribution of ratings, and the reviewer is then asked to calibrate their own bias. This strategy is tested using paintings instead of conference submissions and Amazon Mechanical Turk workers as reviewers. We conclude that auto-calibration through knowledge is feasible in this situation.

1 Introduction

Every year, academics across many fields submit research papers, the culmination of months of work, to their respective conferences. Academic conferences can be a wonderful place to exchange ideas and garner feedback on research questions. Acceptance the acceptance of these papers to the conference is often viewed as a crucial validation, both of the quality of the paper and of the career of the academic.

Acceptance to these conferences is not easy. The following are the acceptance rates of recent conferences in machine learning and computer vision:

- ICCV 2015: 30.3
- CVPR 2015: 28.4
- ECCV 2014: 26.7
- NIPS 2016: 23.6

Unfortunately, acceptance to these conferences is subjective. In the rubric provided to reviewers, papers are rated on four categories. Technical quality, novelty, potential impact, and clarity/presentation. Reviewers for NIPS 2016 were told to rate the papers based on the following scale:

1. - Low or very low quality
2. - Sub-standard for NIPS
3. - Poster level. Only 30% of submissions should reach this stage or higher
4. - Oral level only 3% of submissions should score this high
5. - Award level, only 0.1% of submissions should achieve this score

2 Motivation

Because there is not ground-truth for the rating of the papers, it is hard to judge the overall review process. That said, there are some discrepancies from the proposed rubric and the final distribution of scores. The Table 1 shows findings from Shah et al., which shows the the rubric is clearly mismatched.

Final project report for course CMU 10709 "Fundamentals of Learning from the Crowd", Fall 2017. Instructor: Nihar B. Shah, TA: Ritesh Noothigattu. The formatting style file is borrowed from the NIPS conference.

Over 10x as many papers received award level 5 than the proposed amount. The rest of the categories were also seriously skewed toward higher rankings. As discussed by [1], this caused a high level of papers to receive passing marks, and the task of sorting these papers fell to the area chairs. By concentrating the decisions to a smaller group, there was a higher rate of subjectivity. Also, rejected papers sometimes received scores similar to accepted papers. The goal of this project is to determine the causes of this positive skew in grading, and to find methods that may help limit this phenomena.

3 Related Literature

The problems in the NIPS 2016 conference procedures were raised by [1]. In their 2017 paper, they outlined the entire conference procedure. They also provided an inside look at many of the statistics of the review process including the number of submissions, number of reviewers, and number of papers submitted. Also useful were the distribution rates of responses to each of the questions, and the distribution categorized by final results of the submission, such as whether the paper was given an oral slot, a poster slot, or rejected. The information and problems presented in this research was referenced heavily in the formulation of our experiments.

When formulating my experiments on Amazon Mechanical Turk and understanding how to best propose the task to workers, I referenced [2]. This research was done in the field of political science. The authors present Amazon Mechanical Turk as a cheap and useful method to collect experimental research. The authors also go into detail about the costs of each of their experiments and the ease of recruitment. They offer statistics such as the number of survey completions per day with various titles and payment schemes. They also investigated the reliability of Amazon Mechanical Turk responses.

4 Problem Statement

When rating paper submissions to NIPS, reviewers have biased responses. The ratings of the submissions tend to be much higher than they should be. This causes extra work, as all submissions rated 3 and higher need to be reviewed by a conference area head.

During NIPS 2016, 56% of papers reviewed received a score of three or higher. The desired distribution would dictate that only 30% of papers receive a score of 3 or higher.

Biases are often inherent in rating systems and surveys. While it is possible for responses to be calibrated post-collection, these types of ratings and surveys would benefit from having responses that are already calibrated.

This problem has some key characteristics that need to be realized in any experiments attempting to simulate it. Most importantly, the reviews are subjective. The desired information gained from the reviews is not a true value, but is dependent on the reviews themselves. Also important is that the reviewers have an understanding of the field. NIPS papers are scored on a scale which compares them to all the other submissions. Reviewers must then have a general idea of the average quality of submissions.

5 Experiments

5.1 Design

When choosing an experiment, we needed to choose a situation that would match the characteristics of the NIPS paper review process as closely as possible.

For the experiments, the subject material was chosen to be paintings by famous painters. This domain was chosen to follow closely the criteria outlined above:

- There should be no ground truth. Testing bias and bias calibration with respect to an item that has a ground truth is a different experiment entirely. The goal is to test the bias can be calibrated without a ground truth. As they say, "beauty is in the eye of the beholder." With a domain such as paintings, there is no one who can objectively say that one painting is better than another painting, and there may be a difference between people based on there

personal preferences and background. This is similar to the domain we are trying to emulate, conference papers.

- In the conference submission review, the instructions say to rate submissions based on a reference from one submissions quality against all other submissions. For example, to rate something a 5, the reviewer is supposed to believe that it is in the top 0.1% of submissions. It is not immediately obvious that the domain of paintings satisfy this criteria. Should a reviewer compare works of van Gogh to finger paintings by their friends two-year-old daughter? The solution is to briefly show the viewer a collection of paintings. Enough for them to establish a mental bank of the paintings that they should compare a specific painting to.
- Finally, the domain of paintings suits itself well for use on Amazon Mechanical Turk. Reviewers do not have to be experts on the judgement of paintings, they only need to know where their own preferences lie.

5.2 Procedure

The subjects of the experiment are recruited using Amazon Mechanical Turk's online platform. While they are on the platform, they can preview the survey and see what the reward would be if completed. After some experimentation, it was decided that charging $0.75 for reviewing a subset of 10 paintings was appropriate. If the subjects decided that they were interested in taking the survey for the reward, they would enter and begin the survey.

There were two groups in the experiment. First, the experiment was conducted with group A and the results were gathered. Afterwards, the experiment was ran with group B.

Subjects in group A were then shown a collage of images 1 followed by a series of instructions. These instructions outlined the desired rating scale. The instructions stated that they should rate the paintings on the scale where 5 means the painting is in the top 0.1% of that category, 4 the top 3%, 3 the top 30%, 2 sub-standard and 1 low quality. Because there were only 30 paintings, rating something as being in the "top 0.1%" is ambiguous. The decision of what this meant and when to rate something a 5 was left up to the reviewer.

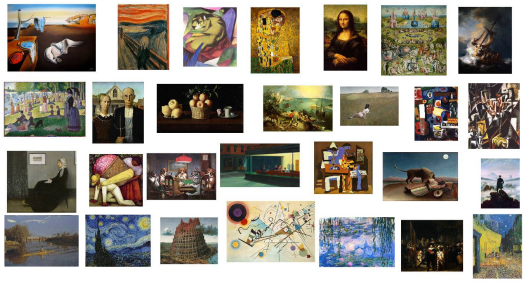

Figure 1: collage of images shown to subjects

Subjects in group B were shown the same instructions, with some additions. First, the phenomena of raters rating too high based on the desired rating system was explained to them. Then, a chart was shown with the previous distribution of responses vs the desired distribution of responses **??**. Finally, they were asked a question to determine if they had understood the explanation of the bias.

After the instructions above, subjects were asked to ranking a subset of 10 paintings from the 30 they were shown in the collage. All subjects in both groups ranked the same paintings in the same order. They were asked to rate the paintings based on their use of color, artist's creativity, and overall quality. The results were then collected by the Amazon Mechanical Turk system and stored in a .csv file.

5.3 Results

Overall, 1800 responses were collected from the Amazon Mechanical Turk workers. These responses were evenly split amongst groups A and B.

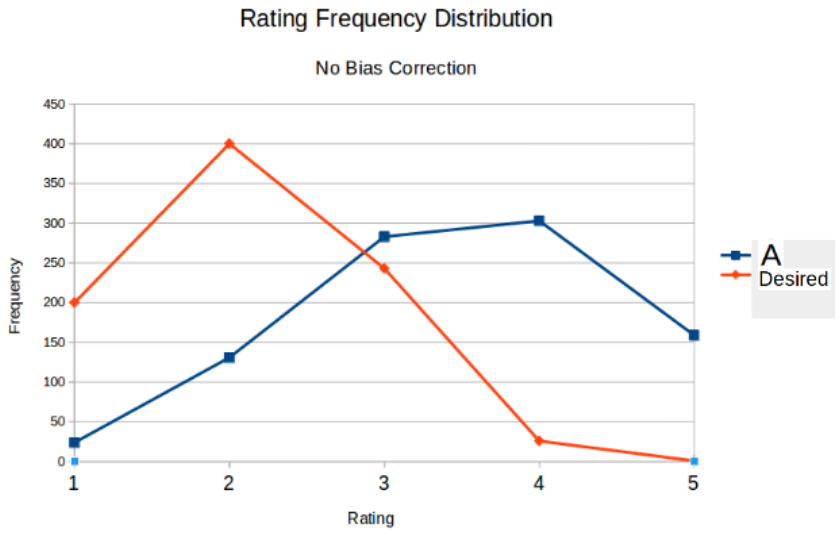

Figure 2: Distribution of group A shown to group B

The final distributions of the three categories, group A, group B, and the desired distribution is shown in **??**. The table showing the scaled % values of the distribution is shown in 4

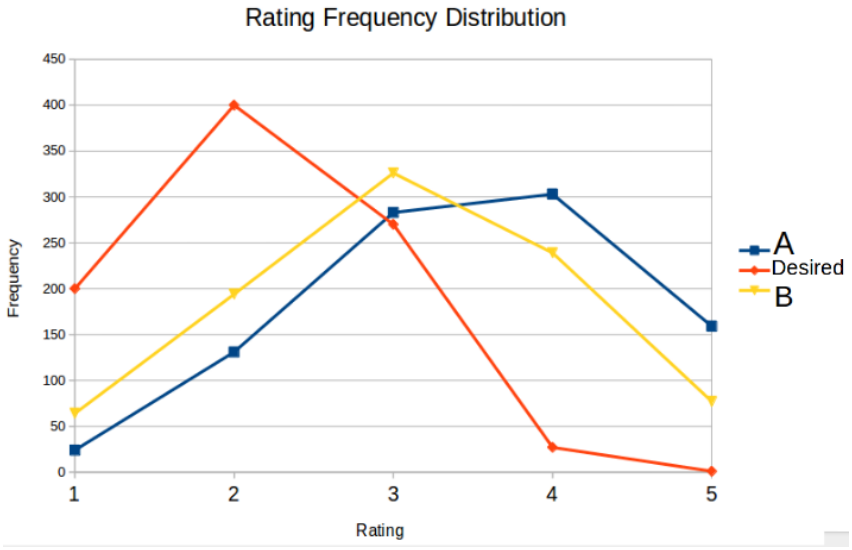

Figure 3: Final distributions from each category

As you can see, the auto-calibration works to some extent, but is very far from completely correcting the problem.

As mentioned before, subjects in group B were asked a question. They were asked "which of the ratings should appear the least often". Exactly 50% of them responded correctly, implying that the other 50% either did not read the instructions or did not understand the distributions. One thing we can do is filter the responses of reviewers in group B by those who answered the understanding question correctly. When this is done, we get the distributions in **??**. These results are much better calibrated.

Rating	A	B	desired
1	2.67%	7.11%	22.22%
2	14.56%	21.56%	44.44%
3	31.44%	36.22%	30.00%
4	33.67%	26.56%	3.00%
5	17.67%	8.56%	0.11%

Figure 4: Final distributions from each category

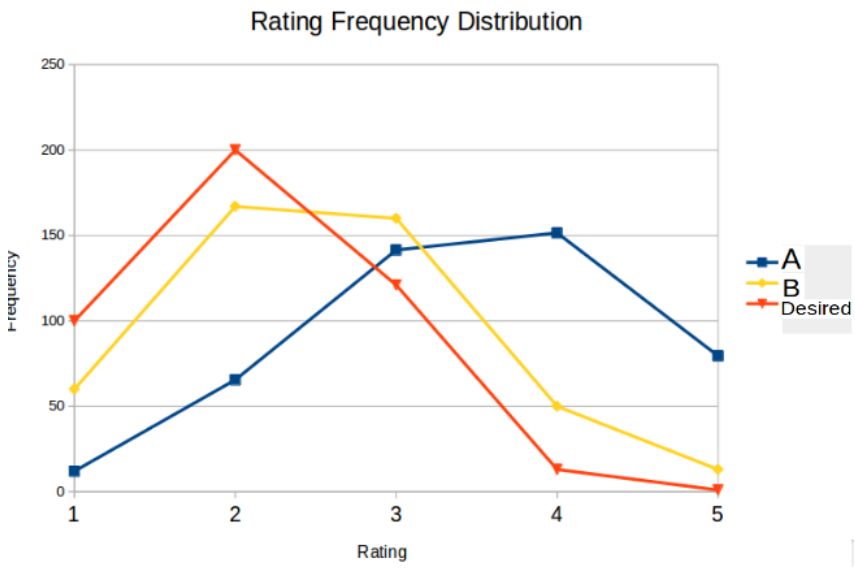

Figure 5: Final scaled distribution with filtering in group B

6 Conclusions

As we see from the results of the experiment, the hypothesis was proven correct in this situation. However, more work is need to be done in order to determine if this is a feasible strategy going forward.

First, the unreliability of the Amazon Mechanical Turk workers is unknown. Experienced workers would understand that an experiment like this, in which they are asked opinion, is easy to spam their answers without thinking or following the prompt. One flaw in the experiment that needs to be corrected in the future was that workers in group A were not tested on their understanding of the instructions. As we saw in group B, this had a significant impact on the results.

Looking forward, if no changes are already made to the review process for NIPS 2018 and there were still problems with bias in the NIPS 2017 submissions, I would recommend to the committee that they inform the reviewers of the past distribution and how it was different from their desired distribution of ratings.

References

[1] Shah, N. B. et al (2017) Design and Analysis of the NIPS 2016 Review Process. In: *arXiv preprint*

[2] Berinsky, A. J. et al (2012) Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk In: *Political Analysis (2012) 20:351-368*