Methods for Limiting Positive Skew of Percentage-Based Ranking

Michael Vander Meiden
School of Computer Science
Carnegie Mellon University
mavm@cmu.edu

1 Background

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
- Oral level only 3% of submissions should score this high
- Award level, only 0.1% of submissions should achieve this score

2 Problem

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. Findings from [1] show that the distribution specified by the rubric is clearly mismatched from the reviewers actual distribution.

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 Shah, 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.

Project report for course CMU 10709, Fall 2017. The formatting style file is borrowed from the NIPS conference.

3 Approach

The current plan is to develop a series of experiments to be performed using Amazon Mechanical Turk. The First step is to determine whether the phenomena can be reproduced. Obviously, we can not expect people outside of the machine learning field to review highly NIPS papers, so we must determine a different material which our reviewers can assess the quality of.

Next, we can use the determined medium to perform a series of experiments changing the parameters of the task. For example, do reviewers perform better with percentage values that fit more cleanly into the papers reviewed by the reviewer? The reviewers for NIPS were supposed to select a 5 grade 1/1000 times, but did not review 1000 papers. How might have this affected the quality of their review?

References

[1] Nihar, S.B. & Tabibian, B & Muandet, K & Guyon, I & von Luxburg, U (2017) Design and Analysis of the NIPS 2016 Review Process