CS280r Final Project Report
Voting Rules for Subset Selection

Anna Sophie Hilgard and Nicholas Hoernle

Abstract

Social choice theory, providing tools for making joint decisions in multiagent systems is becoming popular for tasks such as group recommendation systems, information retrieval, and crowdsourcing (Moulin et al., 2016). We pose a scenario where a crowd is asked to select a subset of information points regarding a topic and we show that theoretically optimal voting rules are not necessarily applicable to this setting. Rather, we suggest different voting rules place varying amounts of cognitive load on the participants, introducing noise into the votes of the participants and ultimately confusing the results. Our conclusion is that cognitive load should be included as a modeling variable for voting rules that involve subset selection.

1. Introduction

There are interesting settings where humans collaborate with each other and with computers to generate content, including documents (e.g. Wikipedia) (Hahn et al., 2016; Kittur et al., 2013), taxonomies (Chilton et al., 2013) and summaries (Khosla et al., 2013). In multi-agent systems settings where qualitative content is being created or used it is important to consider the cost of communication. Given a medical setting, Amir et al. (2015) report that study participants could not review necessary information in a timely manner, nullifying the effect of obtaining complete and correct information from others. Hahn et al. (2016) show that crowdsourcing vendors consistently struggle to balance the amount of information needed convey to a worker to equip the worker with the necessary information to excel at his/her job without overburdening the process with too much information. An added complication is that in many settings, the ideal contextual information that is shared is subjective. Therefore, multiple parties have competing interests in having their contributions addressed. One possible solution is to allow for a single contributor or an outside controller to make these subjective decisions. However, experiences with content generators with a strong hierarchical or dictatorial leadership (e.g. Wikipedia Benkler et al.) have shown that the resulting content is often suboptimal from the viewpoint of the whole and heavily skewed to conform to the opinion(s) of the decision-maker(s). Schwartz (2015) stresses that those situations in which group members have different information and the actions of individuals are interdependent are the most critical to be collectively assessed.

Under these conditions, we see a strong case for adopting social budgeting techniques to crowdsource contextual points.

2. Related Work

Boutilier et al. (2015) shows that if we assume adopt a utilitarian framework in which we hope to maximize the satisfaction of all group members, properly chosen voting rules can ensure that we minimize the maximum difference between the optimal possible satisfaction to all members and that selected by the voting rule in expectation (the regret). It is clear that for a dictatorial selection this could be trivially equal to the worst case if the size of the alternative set is larger than two times the size of the set of options to be selected. In particular, we will seek to test the effectiveness of the subset selection algorithm generated by Caragiannis et al. (2017), which approaches the problem as a variation on the maximin rule. In particular, the authors show that it is possible to derive an explicit utility function which maximizes regret while maintaining consistency with the votes. We use the following expression for maximum regret for a subset selection T:

$$\max_{S \in A_k} \sum_{i=1}^{n} \frac{\mathbb{1}[S \succ_{\sigma_i} T]}{\sigma_i(S)} \tag{1}$$

Where $S \succ_{\sigma_i} T$ indicates that there is no alternative in T preferred to every alternative in S given the utility function σ_i, and $\sigma_i(S)$ is the ordinal ranking of the best alternative in set S in the ranking determined by the utility function σ_i. Intuitively, any term in this maximization captures the lost satisfaction to the voters of not having the given set S_i chosen rather than T, weighted by how much he or she liked his or her best option in S_i. This will lead to a greater penalization for sets T that do not give many participants at least one of their top choices.

We seek the set T that minimizes this quantity.

$$\text{argmin}_{T \in A_k} \max_{S \in A_k} \sum_{i=1}^{n} \frac{\mathbb{1}[S \succ_{\sigma_i} T]}{\sigma_i(S)} \tag{2}$$

Caragiannis et al. (2017) show that this can be solved through an integer linear program (ILP) with nm variables and $nm^2 + \binom{n}{m}$ constraints, where n is the number of voters and m is the number of alternatives available.

For comparison, we also use plurality/knapsack voting, which has been used in real-world participatory budgeting programs (likely in part because of its computational simplicity and ease of understanding Goel et al.) and is shown by Caragiannis et al. (2017) to have empirical regret approaching that of the subset selection algorithm above for subset sizes greater than three, which will be the case in our experiment and should be generally true for problems of this nature.

We apply these different voting rules to the problem of subset selection and conclude that in practice the success of a voting rule may depend heavily on the difficulty (cognitively or subjectively) one has in comparing options.

3. Experiment Design

Three voting rules are compared to evaluate the success of the subset selection. To test the voting rules, we pose a setting where participants are requested to select a number of points that may be relevant to a given topic. We compiled a set of 10 supporting points from popular *New York Times* opinion pieces, and used a web-based survey form to allow participants to make subset selections. The topics were presented in a 'debate prompt' style and the participants were asked to select the points that would contribute the most value to the presented argument. We also allowed participants to provide feedback on the subset selection styles that they most and least enjoyed (Appendix A).

The interface [1] was designed to present participants with three questions (on three different articles) and each question displays a different subset selection protocol. The different selection methods include:

- 'ranking' selection by dragging and dropping alternatives into the correct order from most to least useful for the given argument.
- 'cardinal' selection where each point was given a score out of 10 independently of the others.
- 'plurality' (or knapsack) selection, where check-boxes are selected until 5 selections were made.

The study consisted of 37 respondents and 111 subset choices over the three different voting rules. We solve equation 2 using integer linear programming as described by Caragiannis et al. (2017) to aggregate the ranking and cardinal results into an optimal subset. We induced a ranking over the cardinal results to obtain ranked values to use in the subset selection algorithm, breaking ties

[1] Accessed at: http://nick-and-sophie-harvard-cs280r.s3-website-us-east-1.amazonaws.com/index.html

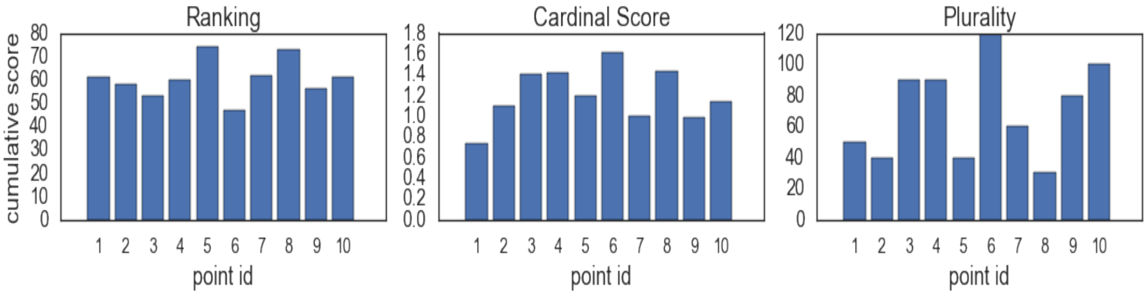

Figure 1: Bar chart of the score each answer point receives under the different voting rules.

at random. For the plurality results, we selected the subset greedily using a majority rule approach. Selected subsets included four points each.

Finally, selected subsets were presented to a different group of 15 participants. These participants were simply asked to select the most relevant subset, given the same topic. This data was also collected through a web-based survey [2].

4. Results

The results here, while indicative of trends are mostly not statistically significant and thus we do not make *significant* claims about our conclusions. Rather, we use the results to inform interesting questions and to pose further work in this area.

The initial results show a large amount of noise and uncertainty in the selections. Figure 1 shows the score that each answer receives for the third topic [3] in the question set under the different voting rules. [*Some useful reference discuss how this uncertainty is expected*]. Simply from the raw data there does not appear to be a large amount of correlation among the voting formats. Table 1 shows the summary of the selected subsets by the voting aggregation rules explained in 3. Points selected by the various methods showed a generally low degree of overlap. For two of the topics, there was one point present in all three subsets, and for all of the topics, at least two points appeared in multiple subsets. Only one pair of options differed by only a single point.

A different set of voters was asked to vote on qualitatively the best subset. Figure 2 shows that the subset from plurality voting produced the most preferred results in all three topics, although for Topic 1 the Cardinal Score was also tied for the most number of votes. The success of simple plural-

[2] Accessed at: http://nick-and-sophie-harvard-cs280r.s3-website-us-east-1.amazonaws.com

[3] See Appendix B for more details on the specific topics

	Topic 1	Topic 2	Topic 3
Ranking	$[2, 3, 6, 7]$	$[2, 3, 7, 10]$	$[3, 5, 7, 10]$
Cardinal Score	$[1, 2, 7, 8]$	$[4, 7, 8, 10]$	$[3, 4, 6, 8]$
Plurality Selection	$[1, 5, 7, 9]$	$[1, 2, 3, 8]$	$[3, 4, 6, 10]$

Table 1: Table showing the selected subset for the different topics for the different voting rules.

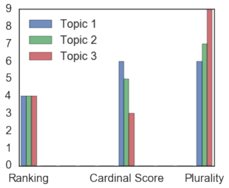

Figure 2: Bar chart showing the final selected subset for each question for each voting rule.

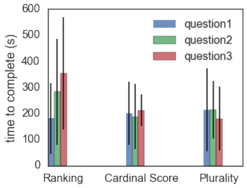

Figure 3: Time to complete the different voting sections.

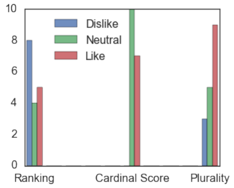

Figure 4: Chart showing the qualitative sentiment towards the different voting rules.

ity voting in our experiment over more complicated minimum regret methods leads to a discussion of the appropriateness of minimum regret in our setting.

The time that it took the participants to complete each voting section was recorded. For Topics 2 and 3, the Ranking vote section took on average 1 minute 30 seconds more time to complete than Cardinal and Plurality voting. However, the results from Topic 1 are inconclusive and the standard deviation on the times is large.

Finally, we turn to the qualitative results that were collected in the initial survey. Here we specifically asked participants to provide feedback on the difficulty of the voting rule, and which voting rule they preferred. An example of this feedback is: *"Ranking is the most difficult. I prefer the format that lets me choose on a scale from 1-10 how strong I think the argument is."* [4]We parsed the qualitative feedback to understand the sentiment towards the different voting rules. When a specific voting format was mentioned positively or negatively, this was recorded with a 1 or -1 respectively. If the voting rule was not mentioned or was mentioned in a neutral setting the score was recorded as 0. Figure 4 demonstrates a trend where 'Ranking' was disliked more than it was liked, 'Plurality' was liked more than it was disliked and 'Cardinal Score' shows a neutral-to-positive sentiment. While the preference for a cardinal over an ordinal method runs counter to generally accepted voting design principles, we believe this may have to do with the number of points that must be considered for any given decision as well as some degree of user frustration at inability to assign ties in the ranking. In particular, to properly rank any adjacent pair of points, the user

[4] All the qualitative feedback is provided in Appendix A

must consider the value of both at once. In assigning a cardinal ranking, the user can focus only on determining the value of a single point.

5. Discussion

The results from Caragiannis et al. (2017) suggest that for a subset of size four, minimum regret should generally provide the lowest upper bound on the regret of participants. That is, it is better in the case where we assume the utility function with highest possible distortion compatible with the selections. However, plurality voting has a very similar upper bound for a subset size of four and may have superior properties in other considerations, as we've seen above. We believe the success of plurality voting in our experiment has to do with two factors.

5.1. Utility Functions and Separability

First, the upper bounds assume the worst case separable utility function compatible with the reported rankings. Even in the case where the utility function is in fact separable, this does not imply anything about the average case, and the two worst case bounds are so similar that it is probable that the average case regret for plurality is in fact better. Furthermore, while the points were selected with the quality that they each could stand alone, we see evidence in the data that respondents may be 'bundling' points, leading to a different type or problem entirely, and one for which the utility function used in minimum regret is a poor representation of the true utility of voters. Plurality, too, is susceptible to voting paradoxes with nonseparable utility as seen in Moulin et al. (2016), but in our case it seems reasonable that the assumption is approximately true (in that some points may be slightly complementary but the addition of a point should never be detrimental in the presence of other points), and so plurality yields a reasonable approximation of the optimal answer.

To investigate the presence of this effect in our collected data, we ran a window of length three over every person's choices, calculating the pairwise distance between all combinations of the three in that window in terms of their ids in our original problem formulation. Because the points were extracted in order from the original opinion pieces (although points are displayed in random order to respondents), numerically close ids correspond to physical and contextual closeness in the original argument. From this, we get an idea of if points that are similar typically get placed with a similar ranking. Figure 5 shows a strong preference for bundling contextually similar points, even though those points are not displayed together in the experiment. In general, Moulin et al. (2016) suggests that in fact there are very few domains in which the separability assumption can be expected to hold. In our experiment in particular, it seems likely that human nature pushes respondents to choose a cohesive argument. That is, one in which the arguments made flow well together. The

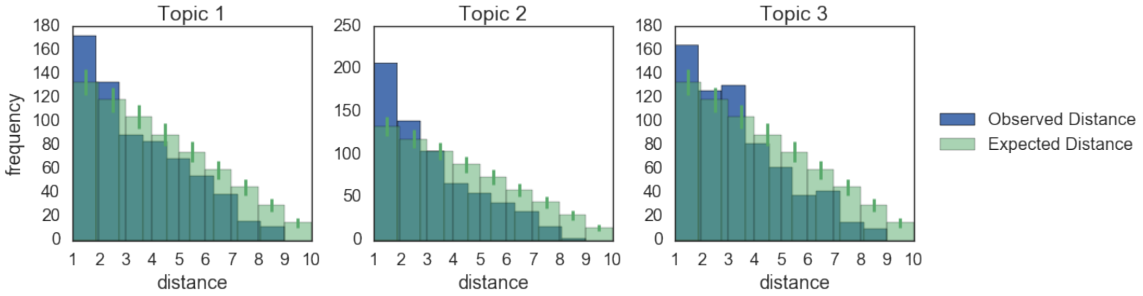

Figure 5: Counts of distances of points in original article when selected in the top or bottom 5 selections.

high frequency of a distance of ‘1’ in the figure 5, corresponding to users selecting points that semantically occurred next to each other in the original article, is significant with a p-value less than [*TODO*].

5.2. Cognitive Load

Second, our qualitative feedback suggests that knapsack/plurality selection is significantly easier for respondents. This leads us to consider the possible effects of cognitive load on the various voting algorithms. In fact, Caragiannis et al. explicitly mention that the analysis has yet to investigate effects of cognitive load although the authors stress in other works, such as Benade et al. (2017) that the entire point of a voting mechanisms is to reduce the unacceptable cognitive load of eliciting a full utility function.

One good reason to do this is that people are likely to make mistakes when presented with a large cognitive load. In particular, in the Sushi dataset from Kamishima (2003), 70% of rankings and ratings of the same subsets contain contradictions. That is, the cardinal values in the rating set do not map to the ordinal values in the ranking set. Then it can be assumed that respondents have reported erroneous preferences in one of the two cases, possibly due to excessive cognitive load of the reporting mechanism. If it can be expected that voters will occasionally make errors in reporting their preferences, we should also be interested in the robustness of these selection algorithms to errors. Previous work has shown that the worst case robustness of both minimax and plurality voting is generally better than many other voting methods when considering the worst case for a single winner and that the worst case for a larger subset is bounded by the worst case for a single winner (Procaccia et al., 2007). Here, we consider the empirical average case for a variety of subset sizes.

To set up the experiment, we take the Sushi dataset mentioned above and calculate for the rating and ranking problems on the same subsets (these are ‘sushi3b.5000.10.order’ and ‘sushi3b.5000.10.score’)

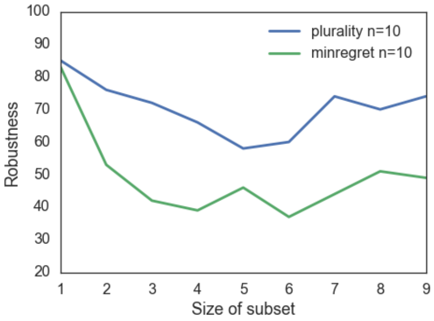

Figure 6: Robustness with 10 voters and 10 choices

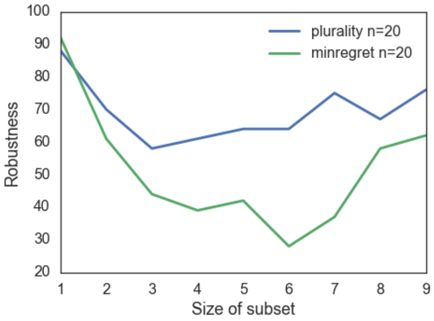

Figure 7: Robustness with 20 voters and 10 choices

the minimum number of flips of adjacent rankings required to bring the ranking into agreement with a ranking induced over the ratings (flips corresponding to equally rated items are not included in the count, as either ranking of such items is consistent with the rating assuming randomized tie breaking). We tally the distribution of the number of errors throughout all 5000 respondents to the sushi survey. Then, we use the third sushi dataset, 'sushi3a.5000.10.order', which contains only 10 types of sushi rather than the 100 in other subsets to bootstrap voting profiles. Because each of the users in the 100 sushi dataset were each only presented with 10 sushi out of the 100, we feel it is fair to assume the same cognitive load would be true for only 10 total sushi types. However, using this dataset allows us to simulate voting over a set of only 10 objects. To create the profiles, we repeatedly draw a set number of voting profiles at random from the rows of the file. We create 100 of these voting profile sets for each trial. Then, we loop through each item in each of the profile sets and induce a number of errors (flips of adjacent rankings) corresponding to a draw from the error distribution. We then perform plurality and minimum regret subset selection on each of the 100 correct voting profile sets and their corresponding profile sets with induced errors and report the number of times the answers matched, in spite of the errors. The results are reported in figures 6 and 7 for subsets of size 1 to 9 and profile sets of 10 and 20 voters each.

We find that in general, plurality voting is much more robust to errors in voting than minimum regret. Then, based on our findings, we would expect that the additional cognitive load of ranking induces more errors in voting and that the algorithm is less robust to these errors.

6. Conclusion

Describes the insights that can be taken away from the work reported in the paper.

7. Future work

Suggests extensions or challenges raised by the project.

H. Moulin, F. Brandt, V. Conitzer, U. Endriss, A. D. Procaccia, J. Lang, Handbook of Computational Social Choice, Cambridge University Press, 2016.

N. Hahn, J. Chang, J. E. Kim, A. Kittur, The Knowledge Accelerator: Big picture thinking in small pieces, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ACM, 2258–2270, 2016.

A. Kittur, J. V. Nickerson, M. Bernstein, E. Gerber, A. Shaw, J. Zimmerman, M. Lease, J. Horton, The future of crowd work, in: Proceedings of the 2013 conference on Computer supported cooperative work, ACM, 1301–1318, 2013.

L. B. Chilton, G. Little, D. Edge, D. S. Weld, J. A. Landay, Cascade: Crowdsourcing taxonomy creation, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 1999–2008, 2013.

A. Khosla, R. Hamid, C.-J. Lin, N. Sundaresan, Large-scale video summarization using web-image priors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2698–2705, 2013.

O. Amir, B. J. Grosz, K. Z. Gajos, S. M. Swenson, L. M. Sanders, From care plans to care coordination: Opportunities for computer support of teamwork in complex healthcare, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, ACM, 1419–1428, 2015.

Y. Benkler, A. Shaw, B. M. Hill, Peer production: A Form of collective Intelligence, Handbook of Collective Intelligence 175.

R. Schwartz, How to Design an Agenda for an Effective Meeting, Harvard Business Review .

C. Boutilier, I. Caragiannis, S. Haber, T. Lu, A. D. Procaccia, O. Sheffet, Optimal social choice functions: A utilitarian view, Artificial Intelligence 227 (2015) 190–213.

I. Caragiannis, S. Nath, A. D. Procaccia, N. Shah, Subset selection via implicit utilitarian voting, Journal of Artificial Intelligence Research 58 (2017) 123–152.

A. Goel, A. K. Krishnaswamy, S. Sakshuwong, T. Aitamurto, Knapsack voting, Collective Intelligence .

G. Benade, S. Nath, A. D. Procaccia, N. Shah, Preference Elicitation For Participatory Budgeting, in: Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI). Forthcoming, 2017.

T. Kamishima, Nantonac collaborative filtering: recommendation based on order responses, in: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 583–588, 2003.

A. D. Procaccia, J. S. Rosenschein, G. A. Kaminka, On the robustness of preference aggregation in noisy environments, in: Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems, ACM, 66, 2007.

Appendix A. Qualitative Feedback from Survey

Appendix B. Question Topics