When to stop the crowd?

JULIAN BERGER, MEHDI MOUSSAID, STEFAN HERZOG , RALPH HERTWIG and RALF KURVERS, Center for Adaptive Rationality, Max Planck Institute for Human Development

1. INTRODUCTION

A powerful approach to increase decision accuracy in high-stake situations is to pool the decisions of multiple decision makers rather than relying on single ones. Arguably, the biggest hurdle for employing such an approach in real life is that experts' time is valuable, leading to a crucial trade-off between accuracy and costs: Adding more decisions typically increases decision accuracy, but at the expense of higher costs. The broad literature describing such a *Wisdom of Crowds effect* has almost exclusively tested the performance of *static* crowd rules (e.g., majority rule, or maximum confidence slating) against the performance of single individuals. A major drawback of such static rules is that all cases are evaluated by the same number of raters, independent of case difficulty, resulting in a relatively high effort. In an ideal scenario, the size of the crowd would depend on the task difficulty, using singletons or small crowds for easy cases, and larger crowds for complex cases. Here we investigate such an approach. We focus on binary-choice tasks, and re-examine eight data sets from the domains of fake news as well as deepfake detection, fingerprint analysis, scientific replication forecasting, geopolitical forecasting, and skin cancer and breast cancer diagnostics.

1.1 Dynamic Decision Rules

We compare the performance of three dynamic stopping rules that rely on confidence judgements against two widely-used aggregation rules: maximum confidence slating (in groups of two) (MCS) (Koriat, 2012), and majority rule (in groups of three) (Hastie and Kameda, 2005). We ask whether and when we can achieve a similar performance to these benchmarks using less than two or three decision makers respectively. We restrict our rules to sample a maximum of three independent decision makers to resemble real-world situations in which collecting additional judgements is costly. Our rules were inspired by fast-and-frugal decision trees and evidence accumulation models. For space reasons, we present only one of the rules here: the Confidence Slating Chain (CSC). CSC uses a confidence threshold t between 0 and 1 and samples individuals until an individual's confidence fulfills t and selects their choice. If t is not met, CSC selects the option associated with the highest confidence. The sampling procedure is as follows: For each case, we randomly draw three individuals and record the performance of the first individual, the MCS using the first two individuals and the majority vote of all three individuals. Next, the three rules are tested. To standardize confidence judgements across data sets, we linearly transform the confidence scales of each data set to match a scaled confidence between 0 and 1; e.g., the skin cancer data set contains a 4-point confidence scale (1-4), a confidence of 1 corresponds to a .25 linear transformation etc. For each rule sample within a case and three random individuals, we vary the threshold t between .1 and 1 with increments of .1, thus increasing the confidence level required reach a decision. Per level of t, we record whether the answer was correct or wrong and the number of individuals sampled. This process is repeated 2500 times for each case within a data set and we report averages across all sampling runs in Figure 1.

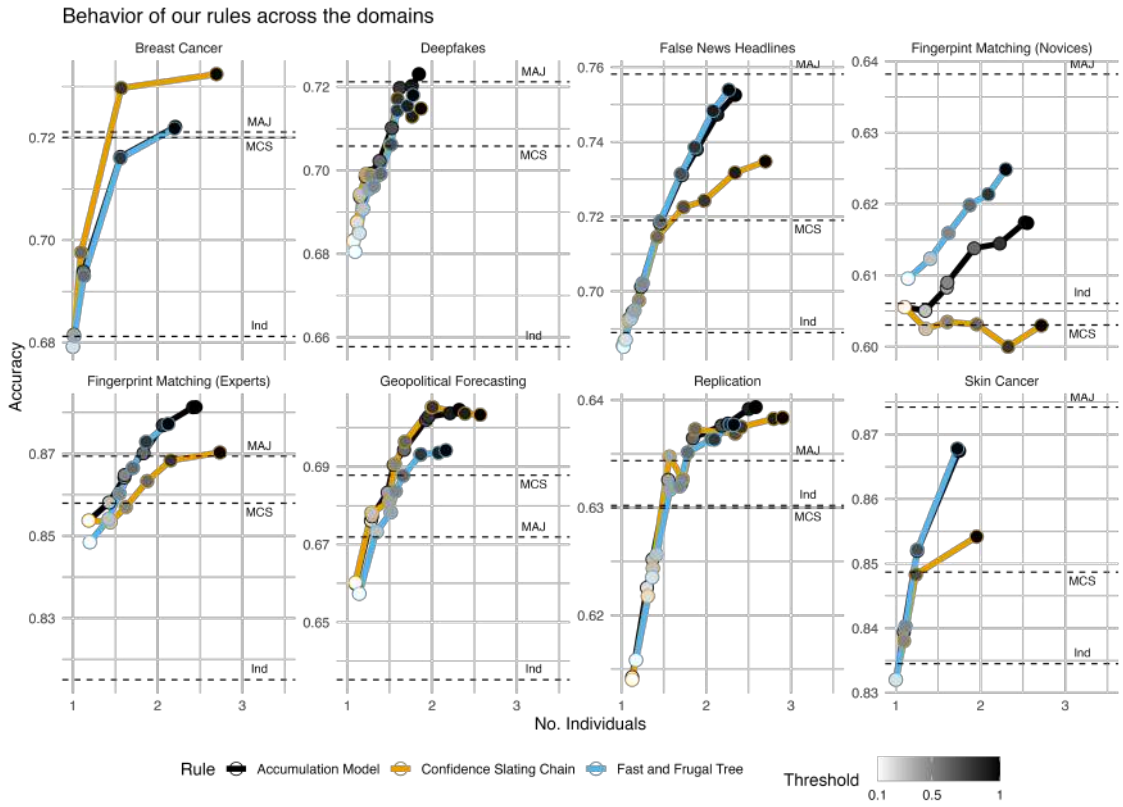

Fig. 1. Performance of the three dynamic rules across the eight data sets. Horizontal lines correspond to the performance of the average individual (Ind), maximum confidence slating (MCS) and the majority rule (MAJ). Colored lines show the performance of the various dynamic rules. Dots on the lines represent the average result of 2500 sampling runs for each rule. Darker colors indicate a higher confidence threshold t, leading to both a higher performance, and a higher number of raters.

2. EMPIRICAL ANALYSES REVEAL REAL-WORLD BENEFITS ACROSS DOMAINS

Figure 1 shows that, as expected, increasing the confidence threshold t (i.e., darker points) leads to higher performance, but also to a higher number of individuals sampled. The horizontal lines indicate the average individual performance, the performance of the MCS in groups of two, and the majority rule in groups of three. The points of interest are where our dynamics rules intersect these lines, as this allows an evaluation of whether a similar performance can be achieved but with potentially a lower number of raters. For MCS, we find that in all eight data sets, the performance of the MCS is met using less than two individuals. For majority rule performance, we observe that in most domains and for most rules, the performance can also be matched with fewer raters, but not in all cases (e.g., skin cancer). However, in five of eight cases, the majority can be outperformed using fewer than three individuals on average and in the cases that this is not possible, our rules were closer than one percentage point to the accuracy of the majority vote.

.

3. NUMERICAL SIMULATIONS: EFFICIENCY OF THE DECISION RULES DEPEND ON CONFIDENCE UTILITY

To explore under which conditions a dynamic decision rule increases efficiency while maintaining effectiveness we use numerical simulations. We simulate environments varying across two axes: accuracy, the probability p of selecting the correct answer; and confidence utility u, the likelihood that a correct answer is associated with a higher confidence than an incorrect one (Moussaid and Yahosseini, 2016). The simulation proceeds as follows: We generate groups of three agents, varying p between 0 and 1 in increments of 0.05. Correct and incorrect answers are each associated with a confidence drawn from one of two beta distributions. The shapes of these beta distributions are systematically changed by creating all possible combinations of the four shape parameters, varying each α and β parameters between 1 and 10. The difference between the means of the beta distributions serves as the confidence utility for each shape parameter combination. For each level of accuracy and combination of shape parameters we simulate 2500 groups and apply the MCS, majority rule, and the dynamic decision rules and record whether the correct answer was chosen as well as the number of sampled agents (fixed values for MCS at two and majority rule at three agents). Figure 2 presents the results of all rules in comparison to the majority rule.

Figure 2A shows the area (blue) in which rules produces a correct answer, which partly rescues wicked cases in which the crowd is worse than chance in selecting the correct answer (i.e., $p < .5$) if confidence utility is high (i.e., $u > .5$). Figure 2B presents areas in which rules are efficient; that is the case when many correct or incorrect decisions get sampled and these are accompanied by high or low utility respectively. To compare the efficiency against the majority rule, Figure 2C highlights areas in which our rules are at least as accurate as the majority rule but uses fewer raters as indicated by the color coding. Figure 2D locates the cases of the eight datasets we examined in the parameter space of the simulations. Comparing Figure 2C and 2D shows, that many of the real-world cases reside in areas in which dynamic stopping rules are at least as accurate as well as more efficient compared to the majority rule.

4. DISCUSSION AND OUTLOOK

Our empirical and analytic results show that under many realistic conditions, using dynamic rules can match the performance of widely-used static aggregation rules using fewer raters. Additional analyses reveal that all rules increase both the sensitivity as well as specificity of decisions in areas in which varying error costs are of importance, such as the cancer detection examples we tested here. We supplement our empirical results further by cross-validating the robustness of our rules with varying confidence thresholds to test how much past data is necessary to predict the efficiency of our rules on new unseen cases. Results indicate that across the eight data sets, even availability of just 10% of the data (e.g., just three cases in the fingerprint data) allowed to accurately predict the accuracy and number of individuals sampled with no statistically reliable error. Altogether, our dynamic decision rules offer new opportunities for making wise use of the wisdom of crowds.

REFERENCES

Reid Hastie and Tatsuya Kameda. 2005. The robust beauty of majority rules in group decisions. *Psychological review* 112, 2 (2005), 494.

Asher Koriat. 2012. When are two heads better than one and why? *Science* 336, 6079 (2012), 360–362.

Mehdi Moussaid and Kyanoush Seyed Yahosseini. 2016. Can simple transmission chains foster collective intelligence in binary-choice tasks? *Plos one* 11, 11 (2016), e0167223.

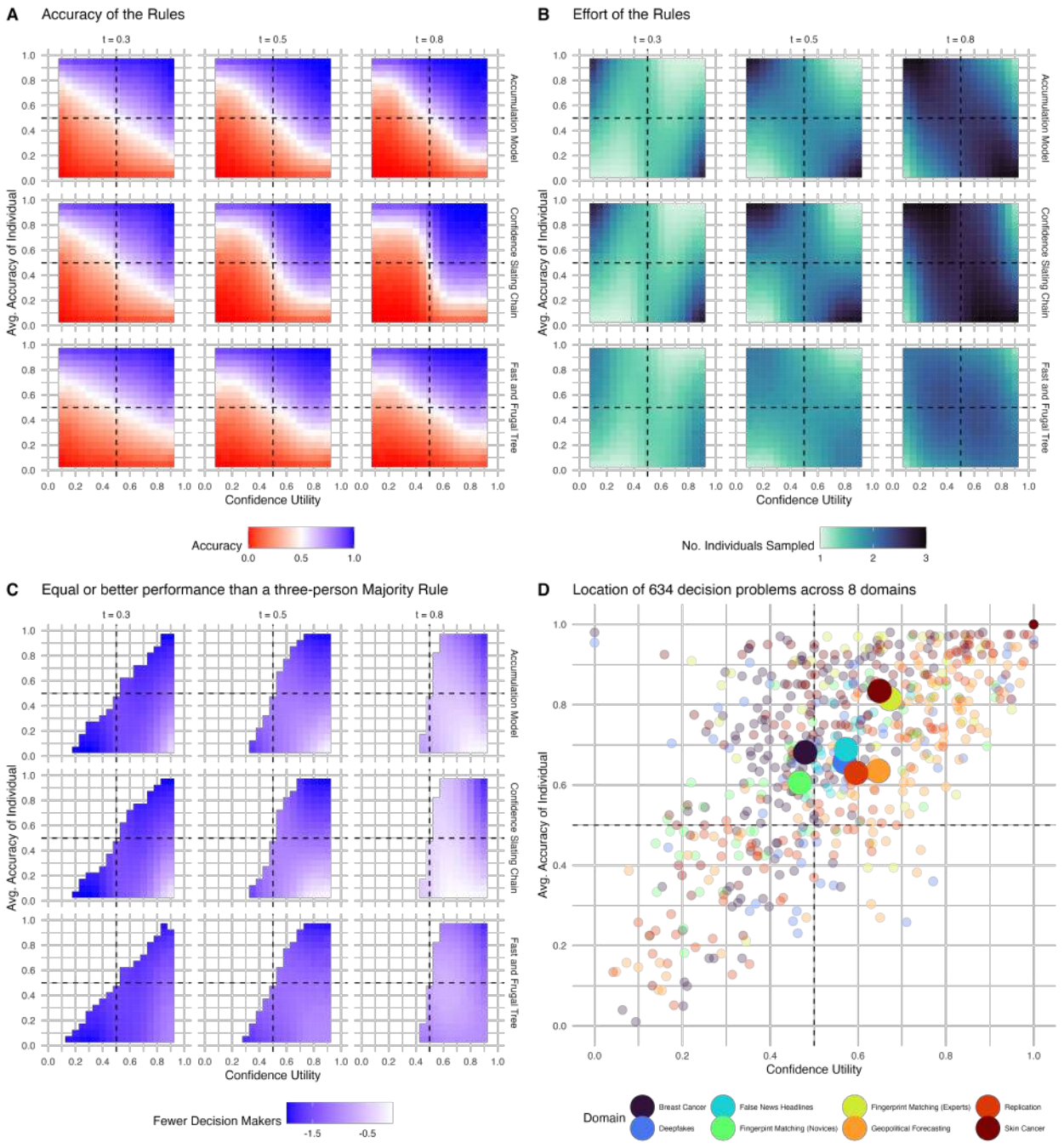

Fig. 2. Simulation results. Confidence utility is the probability that a correct answer is associated with greater confidence than an incorrect one (x axis); accuracy is the likelihood of the average individual choosing the correct solution (y axis). t is the confidence threshold that leads to stopping crowd growth (sub-panel columns in A – C). (A) Expected accuracy of the rules. Increasingly blue (red) colors indicate increasing (decreasing) accuracy of the rules. (B) Expected effort of the rules. Darker colors indicate areas of smaller efficiency in which the rule samples the maximum amount of three agents. (C) Efficiency benefit of the rules over the majority rule. Colored tiles show areas in which the expected performance matches or outperforms the majority rule. Darker colors indicate higher efficiency benefits. A value of -1 indicates that the rules can do at least as good as the majority rule but with, on average, 1 rater less. (D) Locations of the eight data sets in the confidence utility and accuracy parameter space. Small dots represent single cases while large dots present the average values within a data set.