PAPER Special Section on Information and Communication Technology to Support Hyperconnectivity

Learning from Repeated Trials without Feedback: Can Collective Intelligence Outperform the Best Members?

Yoshiko ARIMA[†a)], ***Nonmember***

SUMMARY Both group process studies and collective intelligence studies are concerned with "which of the crowds and the best members perform better." This can be seen as a matter of democracy versus dictatorship. Having evidence of the growth potential of crowds and experts can be useful in making correct predictions and can benefit humanity. In the collective intelligence experimental paradigm, experts' or best members ability is compared with the accuracy of the crowd average. In this research (n = 620), using repeated trials of simple tasks, we compare the correct answer of a class average (index of collective intelligence) and the best member (the one whose answer was closest to the correct answer). The results indicated that, for the cognition task, collective intelligence improved to the level of the best member through repeated trials without feedback; however, it depended on the ability of the best members for the prediction task. The present study suggested that best members' superiority over crowds for the prediction task on the premise of being free from social influence. However, machine learning results suggests that the best members among us cannot be easily found beforehand because they appear through repeated trials.

key words: *collective intelligence, crowd-within effect, cognition task, prediction task, repeated trials*

1. Introduction

Humanity has a long history of learning from and overcoming the adverse effects of war, poverty, or dictatorships through long struggles. However, history is repeating itself, and the same tragedies are occurring again. In predicting events that rarely occur, we seem to repeatedly fail by making naive predictions as if we have forgotten our past experiences. Based on the sad premise that we are not capable of learning from feedback, we examine how far we can go to become smarter.

This study addresses iterative learning without feedback.

The mental models formed by repeated learning may not necessarily lead to correct answers, but they will form an understanding of the framework (familiarity with the problem format and procedures). Under these conditions, the extent to which collective intelligence (average of a cloud) can become smarter will be examined in a comparison between collective intelligence (average) and the best members (highest score). Hereafter in this manuscript, Collective Intelligence will be abbreviated as CI.

In recent years, the concept of CI has attracted attention in various fields, including computer science, psychology, and social sciences. This concept revolves around the idea that problem solving by a group or crowd can result in decisions that are better than those made by experts. For example, in the case of internet searches, mechanical computation of human choices yields better answers than the creation of an expert system. CI is a promising approach to tackling complex problems and has the potential to contribute to society in many ways.

The development of a group intelligence test by an MIT research team was the first step in this research [1], [2]. A general intelligence factor for the group was identified and referred to as "C-factor" in the factor analysis of group performance on tasks. Although reanalysis with hierarchical linear models clarified that the explanatory power of the C-factor was inferior to the expected result [3], some evidence showed that the C-factor predicts CI performance better than individuals' G (general) intelligence scores. The tasks in the MIT team's study were similar to an IQ test with correct answers.

The most expected task for CI would be a future prediction. In fact, the stock market and social networking sites themselves are examples of CI. As far as the bubbles and flames they bring about, they are more closely tied to the madness of crowds than the wisdom of crowds. As long as CI is an accumulation of human judgment, it will be difficult to learn from rare events. On the other hand, crowds have been found to be more accurate in predicting stock prices [4] or reputations as measured by the number of "likes" on Facebook [5]. Therefore, CI might learn some aspects from repeated experiences.

Several studies have shown that the best member predictions were superior to crowd predictions. At NASA, researchers utilized a tournament-type experiment to improve the efficiency of space station photovoltaic panels. In this study, better solutions than those of the NASA experts were obtained from the best member of the public [6]. Moreover, DARPA has applied CI to future predictions, with participants receiving information first and then discussing forecasts of what was likely to happen within a year [7], [8]. Analysis of the data over four years showed that in predicting specific events, the best members, called "super forecasters," were superior to the crowd. This type of task involved predicting "unexpected events" in the future, for which it is difficult to obtain feedback; on the other hand, the prediction of stock price or a number of "likes" are repeated tasks with delayed feedback.

Manuscript received April 17, 2023.
Manuscript revised July 26, 2023.
Manuscript publicized October 18, 2023.

[†]The author is with Kyoto University of Advanced Science, Kyoto-shi, 615–8577 Japan.
a) E-mail: arima.yoshiko@kuas.ac.jp

DOI: 10.1587/transinf.2023IHP0001

Copyright © 2024 The Institute of Electronics, Information and Communication Engineers

The above-mentioned empirical research related to CI is carried out in various research contexts in response to requests from various organizations. Therefore, it is impossible to compare the performance of CI with that of the best members.

[9] compared the CI (average of a cloud) with the best members (the highest-performing individuals) and expert groups. The expert group was selected from past history in the order of their performance, and five members were considered sufficient. Their simulation results showed that the expert group had the widest adaptability range. To validate the simulation results, an analysis was conducted using 40 sets of psychological research data and 50 sets of economic forecast data. "The results showed that, in 53% of the psychological research data, the expert group performed the best, followed by CI in 35% of the cases. The best members performed best in only 13% of the cases. In the economists' data, the group of experts performed best in 68% of the cases, followed by CI in 30%, and the best member in 2%. In all cases, the herd average was superior to relying on a single expert, and relying on the top five was the best result. This result indicates that CI performs poorly on problems that we realistically encounter because of response distribution bias, but the best members perform even worse. The reason why the best members perform worse than average is that there is a repetition factor in the form of history. Even for tasks that involve some degree of competence, the best members' performance does not last for long.

However, the mental models of real-world expert groups are similar, and high competence often comes with the disadvantage of a lack of diversity. The diversity prediction theorem [10] proves that the collective error (squared error of the arithmetic mean of the crowd) is equal to the individual error (average of squared individual error) minus diversity (average squared distance from the individual to the mean). This theorem implies that improving the ability of individuals and increasing the diversity of a crowd contribute equally to the predictive accuracy of the crowd. A trade-off is likely to occur between the average and the diversity of CI [11] because diversity, including lower performance groups, decreases their average performance. If the shared mental models become distorted due to changing times, etc., the correction by CI will not work. Then, what about the best members who are not under the influence of conformity repeating their answers?

1.1 Collective Intelligence within and Between Effects

[12] regarded the CI obtained from repeated trials by one person as a "crowd within effect" and concluded that it was ineffective as CI. In contrast, [13] found that a "crowd within effect" can be obtained through repeated trials. [14] compared the effect of knowing others' answers with that of repeating for themselves and found that the superiority of the crowd-between effect depends on the content of knowledge required for the tasks. If the task was roughly predictable, the correct answer rate improved through individual repetition.

The crowd-within effect offsets sampling variation in one's mental models, but does not derive different mental models. In other words, a crowd between effects would show superiority for tasks requiring various mental models. Therefore, it can be concluded that whether the CI improves through repeated trials depends on the task.

1.2 Task Factor

Since CI research has been conducted in various fields of study, the tasks and indicators are not identical and are thus difficult to compare. Tasks in CI research should be distinguished by error distribution. One has a statistically independent error, which is expected to be offset by the effects of CI (e.g., [15]). The other type of error has distribution bias.

[16] conducted experiments using two types of tasks: perceptual tasks that involved distinguishing figures from a background of white noise and cognitive tasks that involved predicting the weather from environmental measurements. In the present study, we divided tasks into two categories. A "cognitive task" refers to the perception for ambiguous figures which have a correct answer. The other task is the "prediction task," which does not have a straightforward correct answer, but a correct answer that will be revealed in the near future.

In the present research, we use a simple task, counting numbers of dots, as a measure of cognition and a prediction task of predicting the largest number of a class on the cognition task [17]. This task can control the ambiguity of stimulus with the number of dots. Uncountable trials were arranged to occur more frequently in the second half of the trials. This manipulation examines how the participants become accustomed to rare and unpredictable events. We did not introduce feedback to avoid social influence during the tasks.

1.3 Research Question

The failures in relation to rare events, about which it is challenging to learn from feedback, have been repeated by humanity. How about CI and best members? The purpose of this research is to discover which factors will contribute to improving the learning of crowds through repeated trials without feedback.

Two questions are relevant to this research:

1) Does CI improve performance through repeated trials without feedback?
2) If CI improves through repeated tasks, is there any difference between the improvement in cognition tasks and prediction tasks?

2. Method

2.1 Participants

A total of 633 participants (mean age was 19.38; 376 males,

145 females, 29 unknown) participated in this study as students in experiment classes in a psychology course. The number of participants in each class was 8 to 37, and the total number of classes was 32. After deleting missing cases or classes containing different procedures, 27 classes (620 participants) remained.

2.2 Procedure

Ten random dot diagrams were presented to the participants via power point slide projection in the classroom. The number of dots presented varied from 27 to 226, which manipulated the difficulty of this task. Each diagram was presented for 10 seconds. The 27 dots are easy to count in 10 seconds; however, the dot diagrams become ambiguous stimuli over 100 dots. After presenting each diagram for 10 seconds, instructions were given on the slide to ask the participants to individually answer four questions on a printed form distributed on each participant's desk. In this research, one class was one case. Considering the small number of cases, the sequence of 10 slides was not randomized but presented in a fixed order. The number of dots on each slide was 115, 27, 61, 134, 48, 99, 183, 35, 226, and 157, respectively.

After 10 trials were completed, the participants were asked to form four-person groups and make group decisions for each of the 10 trials of the cognition task. After all discussion groups made their decisions, the experimenter gave feedback on correct answers and explained the purpose of this experiment.

2.3 Dependent Variables

1) Q1. Cognition of the number of dots on each slide (Cognition task)
2) Q2. Confidence in the Q1 answer
3) Q3. Prediction of the smallest number of Q1 in their class
4) Q4. Prediction of the largest number of Q1 in their class (Prediction task)

Because the smallest number was limited by the answer given for Question 1, in this study, Question 4, the prediction of the largest number, was analyzed as an open-end prediction task.

3. Results

3.1 Means and Deviations of Each Answer in Each Trial

As shown in Fig. 1, the larger the number of dots is, the wider the range of smallest to largest number predictions. Figure 2 also shows that the SDs increase as the number of dots on each slide increases. The range of predictions of the largest number tended to be larger for the more difficult tasks (slides with more than 100 dots) in the second half of the trials. Accuracy for the easy tasks (slides with fewer than 50 dots) did not change over 10 trials.

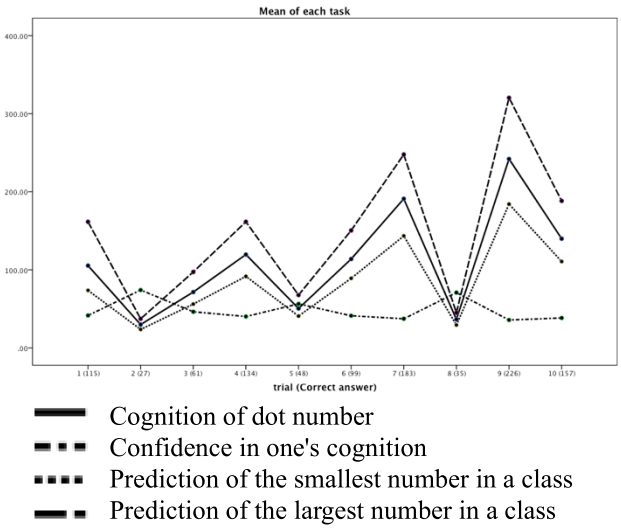


Fig. 1 Mean of answer on each trial

Note. Figure 1 represents the mean of the answers of the four questions on each trial. The X-axis shows the number of dots on each slide (correct answer of cognition task) in the bracket at each trial. Each line represents the mean values of dependent variables as below.

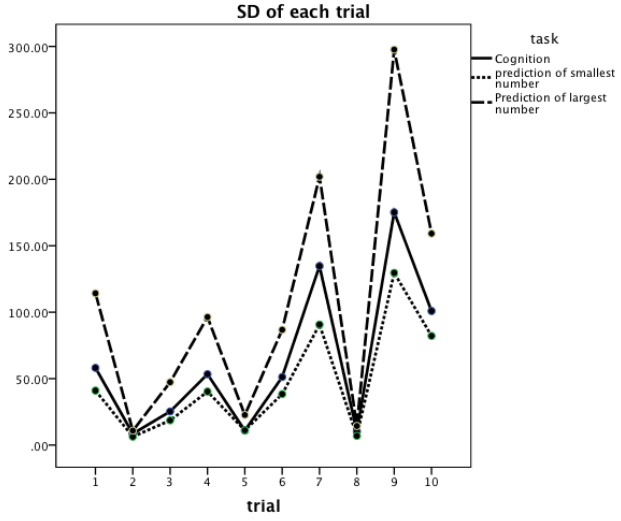


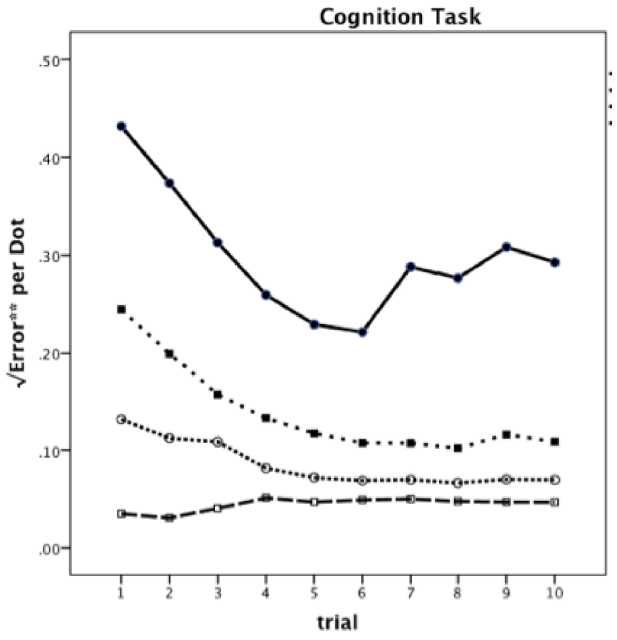
Fig. 2 SD of answer on each trial

3.2 Cognition Task Learning Through Repeated Trials

The error score was calculated using the root square of each answer minus the correct answer. For the cognition task, the correct answer is the number of dots presented on each slide. For the prediction task, the correct answer is the largest number of each cognition trial in each class.

There are four indices for the cognition task.

1) The "individual" index was the mean error score of each individual.
2) The "group decision" index was the mean error score between the group decision and the correct answer.
3) The "CI" index was obtained by calculating the average



Individual: mean of individuals' errors.
Group Decision: error of 4-member groups' decisions.
CI: error of the mean of a class.
Best Member: the error of the best member in a class.

Fig. 3 Mean squared error per dot on the cognition task

Note. The Y-axis represents the cumulative error score divided by the cumulative number of dots.

answer of a class and then calculating the mean error from the correct answer.

4) The "best member" index was the lowest error score until the trial, which was not always the score of the same person across the 10 trials. Fixing the best member at the first selection would increase the error due to iterations and thus the possibility that CI will prevail, but in this study, the best performance at that point is always used as the best member indicator.

For the purpose of representing the learning rate through repeated trials, the error scores were summed up until each trial and divided by the cumulative dot numbers.

3.2.1 Cognition Task

Figure 3 depicts the cumulative error score (sum of error score until the trial) divided by the cumulative dot number, showing that the individual index improved from Trial 1 to Trial 6, and after that, the error score increased.

The increase in error scores, indicating a decline in performance, is consistent with the increase in SD after the 7th trial, as shown in Fig. 2. However, CI improved in performance through the whole sequence of 10 trials; likewise, the best members' performance improved across the 10 trials. This superiority of CI was caused by the distribution of answers, which was symmetrical around the correct answer. Group decisions that were conducted after 10 trials showed a similar pattern of CI but a lower level of accuracy than the CI

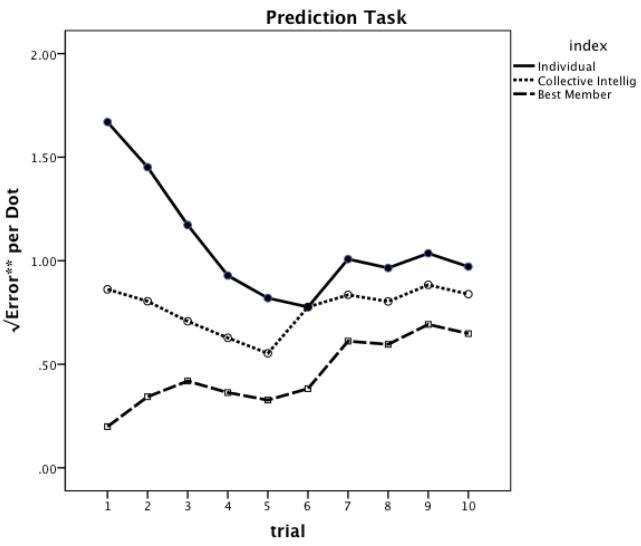


Fig. 4 Mean squared error per dot on the prediction task

Note. The Y-axis represents the cumulative error score divided by the cumulative number of dots.

of a class. The group decision procedure was not conducted for all classes; therefore, the next ANOVA was conducted for the three indices: individual, CI, and the best member for the averaged data of each class (n = 27).

Repeated measure (3 indices X 10 trials) ANOVA yielded the main effect of index, $f(1.09, 17.46*) = 34.30$ $p < .001$ *partial* $\eta 2 = .68$, and the main effect of trial, $f(1.18, 18.84) = 5.93$ $p = .021$ *partial* $\eta 2 = .27$.

3.2.2 Prediction Task

The group decision procedure was only used for the cognition task, so there were three indices, "individual," "CI," and "best member" for the prediction task.

Figure 4 represents the cumulative error scores of the prediction tasks per cumulative dot number.

On the prediction task, the performance of all three indices did not improve after the 5th trial. Repeated measure (3 indices X 10 trials) ANOVA yielded the main effect of index, $f(1.13, 28.35) = 18.25$ $p < .001$ *partial* $\eta 2 = .42$, and interaction of index X trial, $f(1.28, 32.03) = 7.04$ $p = .008$ *partial* $\eta 2 = .22$.

Adding task factor (Cognition or Prediction) to this analysis, repeated measure (2 tasks X 3 indices X 10 trials) ANOVA yielded the main effect of task $(1, 26) = 20.39$, $p < .001$, partial $\eta 2 = .44$, interaction of task X index, $f(1.17, 30.36) = 7.99$, $p = .006$, partial $\eta 2 = .24$, and three-way interaction of task X index X trial, $f(1.36, 35.36) = 4.95$, $p = .023$, partial $\eta 2 = .16$, in addition to other effects found in the above analysis. There is a difference in performance between cognitive and predictive tasks, where index and trial factors had interactions.

Compared with the cognition task, CI for the prediction task could not be learned through repeated trials. Performance drops off in the second half of the trials where trials

became difficult.

3.3 Machine Learning Analysis

To investigate the crucial factor in the learning of CI, exploratory analysis with random forest was conducted for the individual (620 participants) dataset. Random forest is a machine learning technique suitable for time-series prediction, which is often used in machine learning competitions. [18] compared the performance of machine learning competitions using prediction accuracy as an indicator and found that Random Forest had the highest accuracy among the machine learning methods. In this analysis, the error score of each trial was not cumulated. One of the goals of machine learning analysis was to determine when the best members contributing to the final collective knowledge occur (are they competent from the beginning, do they learn, or are they unstable until the end)? As the cumulative value would cancel out the performance of different best members, the value for each trial was used as the feature value. We also decided that using cumulative values was inappropriate because it would lead to autocorrelation among the features. Therefore, the best member calculated here is a different person chosen for each trial, and, in principle, collective intelligence cannot reach above the best member. To normalize the variables, the error score was the root squared percentage error, which means that the difference between each answer and the correct answer was divided by the correct answer and then root squared. The 620 datasets were divided into a training dataset (332 participants, 14 classes) and a test dataset (288 participants, 13 classes). The target of the machine learning was the error score of CI on the cognition or prediction task on the 10th trial. The features (variables) were individual, CI, and best member's cognition and prediction error scores on each of the 1st to 9th trials and group size (n of each class). Although features had correlations among them, multilinearity can be avoided with forest tree analysis. The decision tree regressor and random forest regressor in the scikit-learn library of Python were adapted for model fitting. The number of estimators was 100, the depth of the tree was 3, and the model fitness criteria were *RMSE* (root mean squared error) and *r2* (R-square). Using the decision-tree regressor, training data were used to select ten 10 important features for the target. Using these selected features in a random forest, the model parameters were cross-validated using a grid search. A total of 100 decision trees were computed and the decision tree with the best RMSE was selected. To validate the last decision tree, different data from those used for training and validation were used as test data to verify prediction accuracy.

3.3.1 Forest Tree for the Cognition Task

Figure 5 shows the result of the forest tree targeted at the CI score of the 10th trial on the cognition task. The most recent (9th) error score of CI on the cognition task was crucial. The second factor was the same score as the second trial;

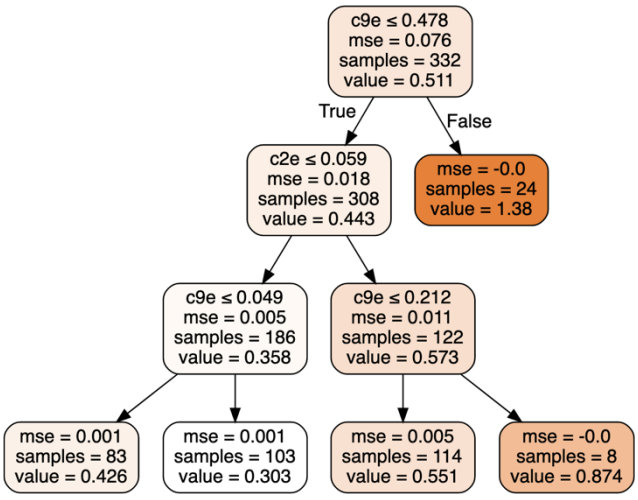


Fig. 5 Forest tree for the cognition task

Note. The target was the error score of CI at the 10th trial.

C9e: The 9th trial's CI error score for cognition
C2e: the 2nd trial's CI error score for cognition
Mse: mean squared error of the model

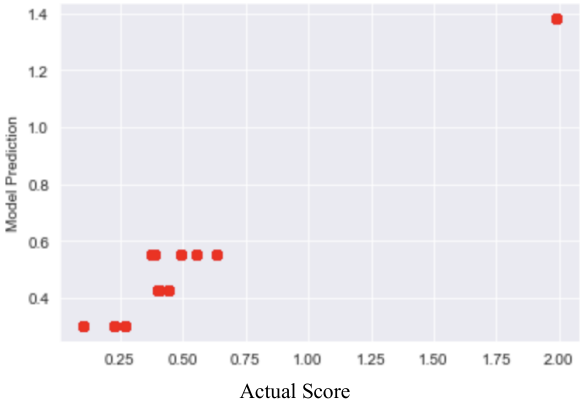


Fig. 6 Fitness of model for the test data of the cognition task

Note. X-axis: Actual error score of the CI at the 10th trial of test data.
Y-axis: Model prediction yielded from the 1st to 9th trial errors of the training data.

however, the 9th trial of the features was the most important (.70), and the importance of the following features was under .09.

This result suggests that CI for the cognition task could learn through repeated trials. When this model was adapted to the test data, the *RMSE* score of this model was 0.2558. There was not much improvement after grid search CV.

Figure 6 represents the model prediction (Y-axis) and the actual score of the test data (X-axis). One dot represents one class. The forest tree model made by the 1st to 9th trial scores of the training dataset yielded an approximation of the 10th error score of the test dataset. The R2 between the actual value of the training dataset and the predicted value of the training dataset was .9702. The R2 between the actual value of the test dataset and the predicted value of the training dataset was .7321. Excluding the outlier (The right upper

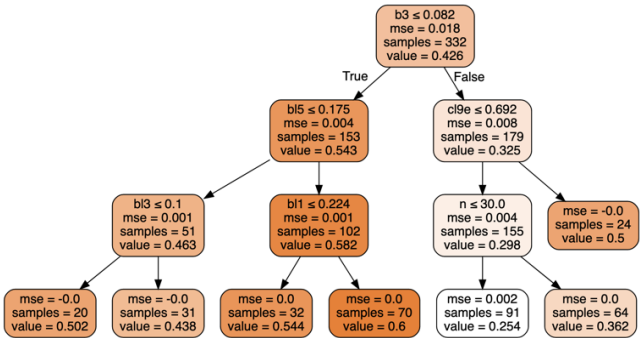


Fig. 7 Forest tree for the prediction task.

Note. The target was the error score of CI at the 10th trial.

B3: The 3rd trial's best member error score for cognition
Bl5: The 5th trial's best member error score for prediction
Cle9: The 9th trial's CI for prediction
Bl3: The 3rd trial's best member error score for prediction
Bl1: The first trial's best member error score for prediction
N: number of each class

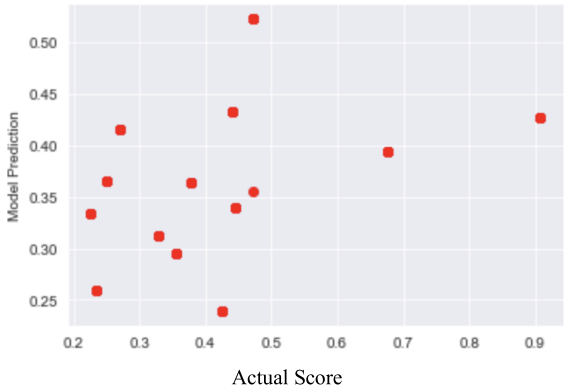


Fig. 8 Fitness of model for the test data of the prediction task

Note. X-Axis: Actual error score of the CI of the 10th prediction task of test data.

Y-Axis: Model prediction yielded from the 1st to 9th trial errors of the training data.

dot in the Fig. 6), the r2 was = −.83 and RMSE was .109.

3.3.2 Forest Tree for the Prediction Task

Figure 7 shows the result of the forest tree targeted at the CI score of the 10th trial on the prediction task. The best member's error score of the 3rd trial on the cognition task was the crucial factor for the CI of the 10th trial. If the first factor was under .082, the second factor was the best member's prediction score of the 5th trial. If the crucial factor was over .082, the CI prediction score of the 9th trial had an effect. The cognitive performance of the best member of the third trial of the features was the most important (.66), and the importance of the following features was under .09.

This result suggests that for the half of the crowds (the better half), who had the best members of better cognitive competence, the CI for the prediction task depends on the ability of the best members. For the half of the crowds (the worse half), where did not have the best members, the recent (9th) error score of CI on the prediction task determined the 10th performance. At the third level of the worse half, the group size (n of each class) affected the performance of the prediction task.

Figure 8 shows the model prediction (Y-axis) and the actual score of the test data (X-axis). One dot represents one class. The random forest model made by the 1st to 9th trial scores of the training data yielded an approximation of the 10th error score of the test data. When this model was adapted to the test data, the *RMSE* score of this model was 0.1989. There was not much improvement after grid search CV. The R2 between the actual value of the training dataset and the predicted value of the training dataset was .9659. The R2 between the actual value of the test dataset and the predicted value of the training dataset was −.0821. The negative values are due to scikit-learn specifications caused by problems with outliers, multicollinearity, nonlinearity, etc.

4. Discussion

For the cognition task, the individual errors increased in the second half of the trials; nonetheless, CI could improve performance. Individuals lose performance in the second half of the trials when the task becomes more difficult, but CI (group mean error) maintains good performance in the second half. Individual mental models improve up to the sixth trial, but after that, the distribution becomes symmetric, suggesting that CI improves. The best members are good in the first three trials, and in the second half, the results are comparable to the CI. For the cognition task, the individual errors increased in the second half of trials; nonetheless, CI could improve performance. Not because of a rise in average ability, the improvement of CI is caused by a symmetrical smoothing of the distribution.

For the prediction task, neither CI nor the best member showed improvement in performance during the second half of the trials. Individuals perform poorly in the first trial but learn rapidly. However, they lose performance in the second half of the trials. CI moved along with individuals. The best members perform best in the first half and gradually lose performance.

For the random forest of regression, the performance of CI for the cognition task was determined by the performance of CI in the previous trials. On the other hand, the model for the prediction task was determined by the best member's abilities. The best member's ability was shown in the 3rd trial. In this experiment, the stimuli of the second trial were so easy that the 1st and 2nd trials seemed to function as practice for the best members. The effect of the best members continued to the last trial for the better half, but it did not influence the worse half. However, because of the poor R2 between the predicted and actual values, the dataset should be reconsidered.

Roughly speaking, if the distribution of answers is symmetrical, such as those of the cognition task, CI will improve

performance through repeated trials without feedback. In this case, CI will eventually reach the best member's performance. However, for the prediction task, CI needs the abilities of its best members in the crowd better than average, who can learn rapidly through trials.

4.1 Limitations of This Study

The difference in machine learning models between the cognition and prediction tasks was apparent; however, this analysis has some limitations. The random forest model for cognition tasks yielded fairly good results, predicting approximately 70% of the test data. For the remaining 30%, this model should be improved by the interaction of variables or other features. The model fitness of the prediction task was worse than the model fitness of prediction; however, the *RMSE* of the prediction model was better than that of the cognition model.

The task in this research was free from social influence. Many topics of CI contain social values. Sharing cognitive schema affects the weighting and evaluation of information, further strengthening bias [19], [20]. This bias is also inevitable in machine learning through crowds as well as expert groups consisting of the best members.

There are concerns that machine learning from big data contains cognitive bias [21]. CI is now being implemented as machine learning. If the correct answer is simple enough to be derived by an algorithm, there is no need for CI. CI is needed for things that require human judgments to determine the correct answer. As long as the feedback of the correct answer is human judgment, recursive machine learning can be distorted.

4.2 Conclusion

CI can improve performance through repeated trials without feedback, especially for cognitive tasks up to the same level of ability as that of the best members. This was not because the average ability of individuals increased but because the distribution became symmetrical. What allows this type of CI to excel is the absence of feedback.

For the prediction task, the conclusions are inconclusive because of the poor R2 between the predicted and actual values. With this premise, this study indicated that the best members have an advantage over the crowd. The best members of the better half of the crowd contributed to the CI's performance. The ability of CI to make predictions may rely on the ability of the best member, the super-forecaster, to learn rapidly.

However, it should be noted that the best members were not the best in the first three trials. In other words, our study suggests that the best members among us cannot be found beforehand.

How can we use our CI to avoid the tragedy of history repeating itself? A direct answer could not be found in this research, but we have a hint for it. We need to assess the experts by their most recent performance, not their reputation. [22] demonstrated that expert groups tend to highly evaluate authorities who had high performance in the past, even if the authority's answers were wrong. Being careful about the unconscious bias hidden in our minds and learning from the best anonymous members will make us wiser and better people.

Acknowledgments

This research was supported by a Grant-in-Aid for Scientific Research (KAKENHI) from the Japan Society for the Promotion of Science (JSPS) (nos. 21K02988 and 22H01072). The author has no known conflicts of interest to disclose. Data supporting the results of this study are available from the corresponding author, Yoshiko Arima, upon reasonable request.

References

[1] A.W. Woolley and L. Aggarwal, "Collective Intelligence in Teams and Organization," Handbook of Collective Intelligence, M.T.M. and B.M.S., Eds., ed London: The MIT Press, 2015.
[2] A.W. Woolley, C.F. Chabris, A. Pentland, N. Hashmi, and T.W. Malone, "Evidence for a collective intelligence factor in the performance of human groups," Science, vol.330, no.6004, pp.686–688, 2010.
[3] M. Credé and G. Howardson, "The structure of group task performance—A second look at "collective intelligence": Comment on Woolley et al. (2010)," Journal of Applied Psychology, vol.102, no.10, pp.1483–1492, 2017.
[4] N.D. Penna, D. Adjodah, and A. Pentland, "Efficiency in Prediction Markets: Evidence from SciCast.org," Collective Intelligence Conference, Santa Clara, Organized by Stanford University, University of Michigan, Facebook, 2015.
[5] S.J. Taylor, A. Peysakhovich, and A.K. Steele, "Predicting Cultural Trends on Social Media using the Crowd," Collective Intelligence Conference, Santa Clara, Organized by Stanford University, University of Michigan, Facebook, 2015.
[6] K. Lakhani, "The Crowd as an Innovation Partner," Collective Intelligence Conference, MIT, Boston, 2014.
[7] J. Matheny, "IARPA's Forecasting Tournaments," Collective Intelligence Conference, 2014.
[8] E. Servan-Schreiver and P. Atansov, "Hypermind vs. Big Data: Collective Intelligence Still Dominates Electoral Forecasting," Collective Intelligence Conference, Santa Clara, Organized by Stanford University, University of Michigan, Facebook, 2015.
[9] A.E. Mannes, J.B. Soll, and R.P. Larrick, "The wisdom of select crowds," Journal of Personality and Social Psychology, vol.107, no.2, pp.276–299, 2014.
[10] S.E. Page, The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies, Princeton, NJ: Princeton University Press, 2008.
[11] C.P. Davis-Stober, D.V. Budescu, J. Dana, and S. Broomell, "When Is a Crowd Wise?," Collective Intelligence Conference, MIT, Boston, 2014.
[12] D. Ariely, W. Tung Au, R.H. Bender, D.V. Budescu, C.B. Dietz, H. Gu, T.S. Wallsten, and G. Zauberman, "The effects of averaging subjective probability estimates between and within judges," Journal of Experimental Psychology: Applied, vol.6, no.2, ed. US: American Psychological Association, pp.130–147, 2000.
[13] K.L. Hourihan and A.S. Benjamin, "Smaller is better (when sampling from the crowd within): Low memory-span individuals benefit more from multiple opportunities for estimation," Journal of Experimental Psychology: Learning, Memory, and Cognition, vol.36, no.4,

pp.1068–1074, 2010.
[14] H. Rauhut and J. Lorenz, "The wisdom of crowds in one mind: How individuals can simulate the knowledge of diverse societies to reach better decisions," Journal of mathematical Psychology, vol.55, no.2, pp.191–197, 2011.
[15] Hong and S.E. Page, "Interpreted and Generated signals," Journal of Economic Theory, vol.144, no.5, pp.2174–2196, 2009.
[16] M.Z. Juni and M.P. Eckstein, "Flexible human collective wisdom," Journal of Experimental Psychology: Human Perception and Performance, vol.41, no.6, pp.1588–1611, 2015.
[17] Y. Arima, Psychology of Group and Collective Intelligence: Springer Nature 2022 (in printing).
[18] J. Sekitani and H. Murakami, "Comparing Accuracy of Time Series Forecasting Methods," The 36th Annual Conference of the Japanese Society for Artificial Intelligence, 2022.
[19] V.B. Hinsz, R.S. Tindale, and D.A. Vollrath, "The emerging conceptualization of groups as information processors," Psychological Bulletin, vol.121, no.1, pp.43–64, 1997.
[20] N.L. Kerr, R.J. MacCoun, and G.P. Kramer, "Bias in judgment: Comparing individuals and groups," Psychological Review, vol.103, no.4, pp.687–719, 1996.
[21] C. O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, New York: Crown Books, 2016.
[22] M.A. Burgman, M. McBride, R. Ashton, A. Speirs-Bridge, L. Flander, B. Wintle, F. Fidler, L. Rumpff, and C. Twardy, "Expert status and performance," PLOS ONE, vol.6, no.7, e22998, 2011.

Yoshiko Arima received her Ph.D. degrees from the School of Human Sciences, Osaka University, Japan, in 2003. She was a technical officer of Osaka University and an associate professor at Poole Gakuin University before taking her present post in 2000. She now with the Center for Social and Psychological Research of Metaverse at Kyoto University of Advanced Science.