機械学習の分類予測に基づく参考回答提示による クラウドワーカの学習効果

松原　正樹†　　小林　正樹††　　森嶋　厚行†

† 筑波大学 図書館情報メディア系　〒 305-8550 茨城県つくば市春日 1-2
†† 筑波大学 大学院図書館情報メディア研究科　〒 305-8550 茨城県つくば市春日 1-2
E-mail: †{masaki,mori}@slis.tsukuba.ac.jp, ††makky@klis.tsukuba.ac.jp

あらまし　本稿はマイクロタスク型クラウドソーシングにおいて機械学習の分類予測に基づいた参考回答提示がタスク結果の品質改善とクラウドワーカの学習効果にどう影響するか検討する。400 名のクラウドワーカを対象に実験を行なった。実験では画像分類タスクを対象に以下の 4 つの参考回答タイプを比較した：正解ラベル、ランダム、正解ラベルによって訓練した分類器の予測、人間の回答によって訓練した分類予測。実験の結果、参考回答の正解率が高い順に応じて品質が改善された。しかし、学習効果に関しては、分類器を用いた参考回答はどちらも予測精度が 100 %でないのにも関わらず、これらの参考回答タイプのみにおいて観察された。言い換えると「正解」や「ランダム」では学習効果が観察されなかった。この結果から次の仮説が示唆された。機械学習は何らかの予測モデルを作るので、人間が機械学習の出力を解釈できる可能性がある。特に問題の完全な解釈が困難な場合は、ランダムな回答だけでなく正しい回答も解釈するのが困難なため、正解を全て提示されるよりも学習しやすい。
キーワード　クラウドソーシング, 機械学習, 不随意学習

Notice

This paper is originally published in IEEE HMData2018 [1].

1 Introduction

Quality assurance is one of the primary issues in crowdsourcing; numerous studies have addressed the problem of ensuring the quality [2]. To ensure the quality, training the workers is the one of typical strategy. For example for the microtasks such as categorization or labeling task, three approaches are commonly used: (1) Showing the experts examples [3], [4], (2) Providing feedback about the worker's performance [5], and (3) Asking the workers interact with each other for performing collaboratively on the same task [6].

As the comparatively simple first approach, Shah and Zhou [7] proposed *self-correction* which is a two-stage setting where the worker first answers the questions and is then allowed to change his / her answers after looking at other workers' answer as a reference. The effectiveness of self-correction has shown theoretically and also empirically [8]: Self-correction improves the quality after the worker answer and induces involuntary learning when the reference was correct answer or answer made by workers who get higher accuracy.

However, when considering the actual operation, the problem is that we need to gather all answers before providing the reference answers. In particular, when an unknown or an urgent problem occurs (*e.g.* natural disaster response, local area problems)— in other words, no experts are around there, no correct answers are gathered, and no investments are made, it is difficult to provide answers as a reference with a high accuracy rate in self-correction to ensure quality.

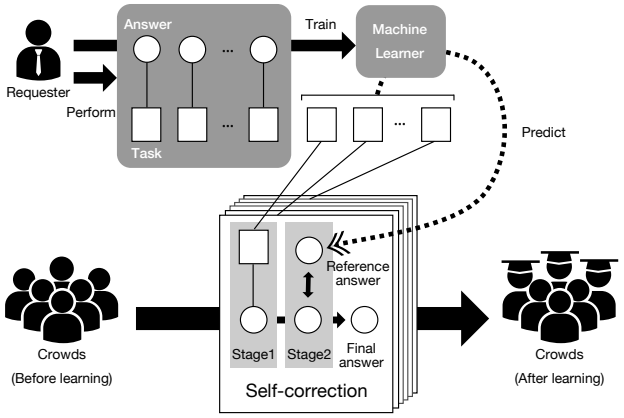

図 1　Proposed method overview: The requester performs the tasks and trains the machine learner based on the answers. After that, the crowds are asked to perform the two-stage setting tasks called self-correction. They answer the task first and are then allowed to change their answers after looking reference answer made by machine prediction. Through performing self-correction tasks repeatedly, crowds are expected to get a learning effect.

To solve this problem, we propose using machine learning prediction as a reference answer (Fig. 1). The requester performs the tasks with the learning data and trains the machine learner based on the answers. Then, the crowds are asked to perform the self-correction tasks and allowed to change their answers after looking

reference answer made by machine prediction. Some of the readers might come up with a question; why don't we use machine prediction as the answer for all the remaining data? The reason is that the accuracy rate is not necessarily high if a machine learner is trained with the answers by the requester instead of the correct answers.

We expect the self-correction with the machine prediction derives better quality as same as self-correction with other worker's answer, but it has not been clarified so far. Therefore, the question arises whether there is effectiveness regarding learning effect and quality assurance in self-correction.

In this paper, we investigate how the task results improve concerning quality during and after presenting machine prediction as a reference answer in self-correction. Four reference types were examined in the experiment; Correct, Random, Machine prediction trained with correct answers (ML-Correct), and that trained with human answers (ML-Human). Our key findings are as follows:

（1） Significant learning effects were observed in "ML-Human" and tendency to that in "ML-Correct", although those accuracy rates were far from correct (100%). Moreover, there were no significant learning effects in "Correct" and "Random". This suggests the following hypothesis: Since machine learners make some "models" for the problem, it is easier for humans to interpret the outputs of machine learners than the results without via them; it is more difficult to interpret not only random answers but also the correct answers in case where the perfect interpretation of the problem (and thus the correct answers) is difficult.

（2） In presenting the machine prediction trained with the human answers, some workers whose accuracy rate is under the machines in the pre-test performed with higher accuracy rate than machines in the post-test. This suggests using the machine prediction can be useful for bootstrapping solutions in the situation where unknown problems occur without expertise or at a low cost.

2 Related Work

Improving the quality of results is an important issue in the crowdsourcing, and numerous studies have addressed these issues [2], [9]. Aggregating the results [10], selecting people [11], and incentivizing people [12] have been common approaches to improve the outputs. Note that our approach can be combined with any of them.

Feedbacking from others or experts is also another common approach to improve the worker outputs. Revolt [13] and Microtalk [14] give workers opportunities to change their answers after seeing justifications of other workers' answers. Shepherd [5] allows both self- and external-assessments of various forms. Self-correction proposed by Shah and Zhou [7] offers a simple extension, which is presenting other workers answer in the task. In this study, due to the simplicity, we adopted self-correction, but our approach doesn't restrict to it.

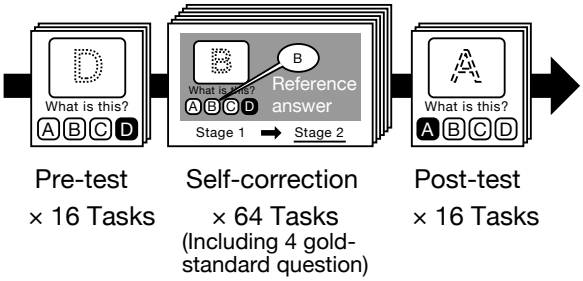

図 2 Experimental procedure

Furthermore, many studies have addressed training people in crowdsourcing (*e.g.* Showing the experts examples [4] and mentoring [15]). Those methods showed the effectiveness of quality improvement; however, they require us to prepare additional training tasks for the workers and to know the answers in advance to teach them to the workers. On the other hand, self-correction does not require any additional task type, and Kobayashi et al. [8] showed that repeating self-correction tasks induce learning effect.

Our approach extends Kobayashi's method; namely, we utilized machine prediction as a reference answer. Since all reference answers were made by machine prediction, we do not have to gather all answers before presenting the references.

As machine learner techniques are recently developed, human-ML interaction for teaching crowds, called *machine teaching* has emerged [3], [16]?[19]. Our contribution is that the experimental results showed machine prediction induced crowds to learn while the correct answers did not work. Applying this finding to machine teaching technology can produce more effective improvement, and further investigation will be part of our future work.

3 Experimental Method

We conducted an experiment to investigate whether presenting machine prediction affects the result in terms of learning effects and quality assurance. This experiment was approved by the local ethical committee.

3.1 Participants

Four hundred workers participated in the experiment via Yahoo! Crowdsourcing[1] and actual tasks were assigned from Crowd4U[2] as the external task. The workers were divided into four groups as each condition mentioned later. The workers were to receive a reward of about $1.00 when they completed all the tasks.

3.2 Procedures

The experiment procedure consists of 3 phases; pre-test, self-correction, and post-test (Fig. 2). Pre-test and post-test were designed for assessment of worker ability. In both tests, workers were

1 : https://crowdsourcing.yahoo.co.jp
2 : https://crowd4u.org

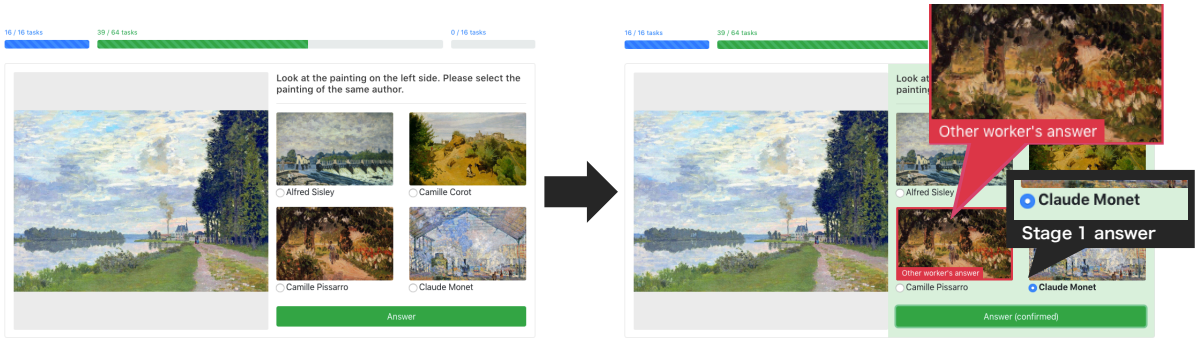

図 3　Self-correction tasks: (Left) In stage 1, a worker answers a question, (Right) then, in stage 2, the worker is allowed to corrects his / her answer by reviewing other workers' answers. In both pre- and post-test workers perform four-class classification tasks that equal to stage 1.

asked to perform 16 tasks. By comparing the accuracy rate between the pre-test and post-test, we can clarify the involuntary learning effects of self-correction.

In the self-correction, the workers were asked to perform 64 self-correction tasks. As we describe in section 1, self-correction is a two-stage setting task which allows the workers to correct their answers by reviewing other workers' answers. By comparing the accuracy rate between stage 1 and 2, we can verify the quality improvement effects of self-correction.

During the 64 self-correction tasks, we set four gold-standard questions to filter out the spam workers. In the gold-standard question, one of the four exemplary paintings was presented to be classified. We analyzed only those workers who correctly answered the gold-standard questions.

3.3 Tasks

In the pre-test, post-test and stage 1 in the self-correction, we used a four-class classification task (Left side in Fig. 3). The classification task involved answering a question by selecting a particular image. We displayed an image of a painting on the left side of a screen. On the right side, we presented four exemplars of painting with painter names underneath. From these choices, workers were asked to identify the paintings on the left.

In stage 2 in the self-correction, we presented the question image and worker's choice from the first stage again. Workers can change their choice. We highlighted other workers' answers as reference answers in the experiment involving self-correction with reference answers (Right side in Fig. 3).

3.4 Conditions

We examined four reference types for self-correction in the experiment; Correct, Random, Machine prediction trained with correct answers (ML-Correct), and that trained with human answers (ML-Human).

Correct: The Correct answers were shown as a reference.

Random: The answers selected randomly were shown as a reference. The expected accuracy equals to chance level which was 25%.

ML-Correct: The predictions of machine learning trained with correct answers were shown as a reference.

ML-Human: The predictions of machine learning trained with human answers were shown as a reference.

表 1　Confusion Matrix of human labeling for learning data

		Predicted			
		Sisley	Corot	Pissarro	Monet
Actual	Sisley	55.6	14.0	13.3	8.0
	Corot	4.3	64.9	20.0	1.8
	Pissarro	28.2	16.7	44.8	17.0
	Monet	11.1	3.5	21.0	72.3

表 2　Accuracy rate before and after filtered out in pre-test

Condition	Filter	N	Median	Mean	Std
Correct	No	93	0.438	0.399	0.135
	Yes	85	0.438	0.396	0.138
Random	No	82	0.438	0.411	0.147
	Yes	69	0.438	0.421	0.148
ML-Correct	No	89	0.375	0.395	0.155
	Yes	82	0.375	0.399	0.152
ML-Human	No	94	0.438	0.408	0.168
	Yes	83	0.438	0.414	0.168

3.5 Datasets

The tasks were to classify the presented paintings by four famous impressionism painters; Alfred Sisley, Camille Corot, Camille Pissarro, and Claude Monet. The painting images were gathered from WikiArt.org[3].

We divided the images into two groups; we used 1200 images as training data for machine learner and used 96 images for experiment tasks.

3.6 Machine Models

We constructed machine learner models with the training datasets

3 : https://www.wikiart.org/

表 3　Confusion matrix of prediction by ML-Correct

		Predicted			
		Sisley	Corot	Pissarro	Monet
Actual	Sisley	73.9	0	26.1	0
	Corot	4.3	87.0	8.7	0
	Pissarro	0	13.0	82.6	4.3
	Monet	17.4	4.3	17.4	60.9

表 4　Confusion matrix of prediction by ML-Human

		Predicted			
		Sisley	Corot	Pissarro	Monet
Actual	Sisley	65.2	17.4	17.4	0
	Corot	4.3	78.3	17.4	0
	Pissarro	34.8	30.4	21.7	13.0
	Monet	30.4	0	4.3	65.2

mentioned above. We used Google AutoML Vision[4] for model construction. Google AutoML Vision is a neural network based machine learning model builder for image recognition, offered as a service from Google Cloud. In this experiment, we made two models; ML-Correct and ML-Human.

ML-Correct: We constructed the machine learner model with pairs of 1200 images and its correct answer as training data.

ML-Human: We asked a person as requester role to label the 444 images to make the training data. We constructed the machine learner model with pairs of 444 images and human answer as training data. Table 1 shows the confusion matrix of human answers for training data. The average accuracy was 59.4%.

4 Results

Pre-test results: We analyzed the number of participants who completed the tasks in each condition and their accuracy before and after filtered out in pre-test (Table 2). Totally 358 out of 400 participants completed the tasks, and 319 participants correctly answered the gold-question tasks. The multiple comparison tests showed that there are no significant differences among conditions.

Accuracy of machine predictions: Table 3 and Table 4 show confusion matrix of ML-Correct and ML-Human respectively. Average accuracy rate of ML-Correct was 76.1% and that of ML-Human was 57.6%.

Learning effects: We analyzed the learning effect of self-correction, namely quality improvements between pre-test and post-test. Fig. 4 shows the accuracy rate between pre-test and post-test in each condition. We conducted a two-way ANOVA with the test phase (pre- and post-test) and the condition (four conditions) as factors. As a result, there was a tendency for the test phase ($F_{(1,315)}$ = 3.565, p = .060), but there was no significant difference in condition ($F_{(3,315)}$ = 1.385, p = .247) and their interaction ($F_{(3,315)}$ = 1.386, p = .247).

4：https://cloud.google.com/vision/automl/docs/

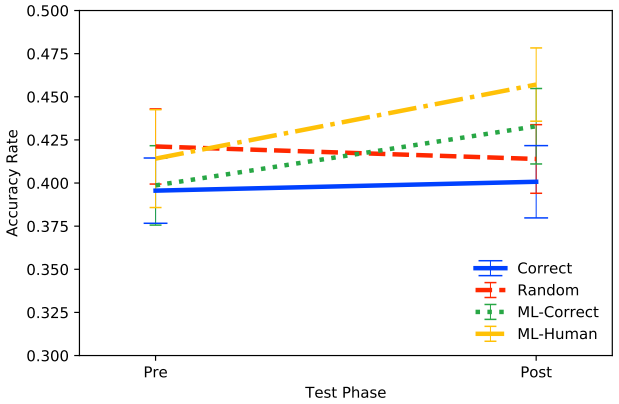

図 4　Learning effects: Accuracy rate between Pre-test and Post-test. Statistical test showed post-test was higher than pre-test in “ML-Human” and “ML-Correct”. In contrast, there were no significant difference in “Correct” and “Random”.

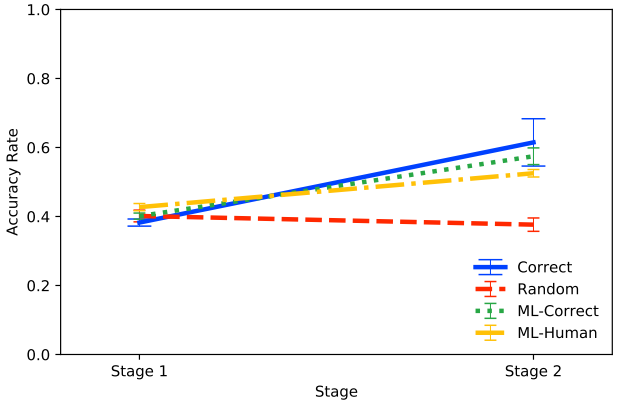

図 5　Quality improvements: Accuracy rate between stage 1 and 2. Statistical showed stage 2 was higher than stage 1 in “Correct”, “ML-Correct”, and “ML-Human”.

Though there was no interaction, we conducted multiple-comparison for each group, since we were interested in the difference between two test phases. A multiple-comparison using t test with Bonferroni correction revealed a significant difference between the pre-test and post-test in “ML-Human” ($F_{(1,315)}$ = 4.88, p = .028), and tendency in “ML-Correct” ($F_{(1,315)}$ = 3.08, p = .080). In contrast, there were no significant difference in “Correct” ($F_{(1,315)}$ = .07, p = .789) and “Random” ($F_{(1,315)}$ = .12, p = .734).

Quality improvements: We analyzed quality improvements during self-correction tasks. Fig. 5 shows the accuracy between stage 1 and 2 in self-correction in each condition. We conducted a two-way ANOVA with the stage and the condition as factors. As a result, there were significant effects from the stage ($F_{1,315)}$ = 183.287, p <.001), from the condition ($F_{(3,315)}$ = 11.957, p <.001), and their interaction ($F_{(3,315)}$ = 37.476, p <.001). Post-hoc t test with Bonferroni correction showed a simple main effect from the stage in the “Correct” ($F_{(1,315)}$ = 185.13, p <.001), “ML-Correct” ($F_{(1,315)}$ = 99.67, p <.001), and “ML-Human” ($F_{(1,315)}$ = 32.28, p <.001), but no main

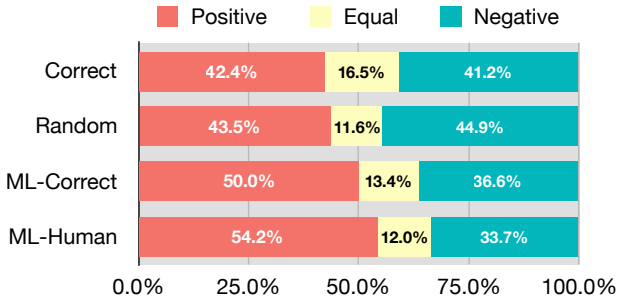

図 6 Proportion of improvements: Difference of accuracy rate between pre- and post-test. Positive indicates post-test is higher than pre-test, and vice versa. More participants in ML-Human improved their accuracy at post-test than those in the others condition did.

effect in the "Random" condition ($F_{(1,315)} = 1.76$, $p = .186$).

5 Discussion

Learning effects: Fig. 4 showed that learning effect by presenting machine prediction was observed. Fig. 6 shows the proportion of difference of accuracy rate between pre- and post-test. Positive indicates post-test is higher than pre-test, and vice versa. As a comparison among the conditions, more participants in ML-Human improved their accuracy at post-test than those in the other conditions did.

As Table 3 and Table 4 showed, the accuracy rates of reference answers in "ML-Correct" and "ML-Human" (76.1% and 57.6% in average) was far from "Correct" (100%). Nevertheless, interestingly, the learning effect was observed in conditions with both machine prediction. In addition, as Fig. 7 showed quality improvement between pre-test and post-test in each class, accuracy rate increased particularly for the "Corrot" class in both "ML-Correct" and "ML-Human", while those didn't much increase in "Correct" in the case where even perfect accuracy references were presented.

These results suggest the following hypothesis: Since machine learners make some "models" for the problem, it is easier for humans to interpret the outputs of machine learners than the results without via them; it is more difficult to interpret not only random answers but also the correct answers in the case where the perfect interpretation of the problem is difficult.

This hypothesis can be explained in terms of *Zone of Proximal Development* (ZPD) by Vygotzky [20]. ZPD is an area of learning that occurs when a person is assisted by a teacher or peer with a higher skill set. ZPD is also defined as an area between what a learner can do without help, and what they cannot do even with help. In this experiment, presenting machine predictions might be suitable for ZPD, in other words, the machine prediction can be useful for scaffolding participants learning [21]. However, presenting the correct answers was out of ZPD, because those reference answers were too complicated to interpret for the participants; the tasks performed in this experiment were difficult. Since the results would be different in easier tasks, further investigations with other datasets can be future work.

The results also showed that approximately 10% workers whose accuracy rate is under the machines at first achieved higher accuracy rate than machines after learning. This implies that this strategy can be expected to be useful for bootstrapping solutions in the situation where unknown or urgent problems occur without expertise or at a low cost.

Quality improvements in each class: Fig. 5 showed that quality improvements by self-correction were observed in "Correct", "ML-Correct", and "ML-Human". Comparing Fig. 4 with Fig. 5, it is interesting that the order of effectiveness were not same between learning effect and self-correction effect; "ML-Human", "ML-Correct", "Correct" and "Random" in learning effect whereas "Correct", "ML-Correct", "ML-Human", and "Random" in self-correction effect.

Looking into each class, as Fig. 8 shows, accuracy rate seems to be higher according to the accuracy rate of reference answers. Presenting the higher accuracy reference answers gives the larger difference between stage 1 and 2; the accuracy rates of reference answers with all classes in "Correct", "ML-Correct", and all classes except "Pissarro" in "ML-Human" were over 60% and thus improvements appeared in those groups. Analyzing the answer pattern, this was because some participants often follow the reference answers. Since the prediction accuracy was 21.7% with "Pissarro" in "ML-Human", the quality decreased from stage 1 to 2 in that group. Nevertheless, the accuracy rate in "Random" remained almost the same instead of decreasing due to the low reference accuracy because participants did not follow the reference answers.

6 Conclusion

In this paper, we investigated whether presenting machine prediction affects the result in terms of learning effects and quality assurance. Learning effects appeared only in presenting machine prediction, although the accuracy rate of machine prediction was far from correct. This suggests the hypothesis that it is easier for humans to interpret the outputs of machine learners than the results without via them, such as the correct answers for complicated problems.

The results also showed that some workers achieved a higher accuracy rate than machines after learning. This implies that this strategy can be expected to be useful for bootstrapping solutions in a situation where unknown or urgent problems occur without expertise or at a low cost. Combination of selecting high-quality workers [22] and iterative active learning methods [18], [23], [24] may help the bootstrapping.

For future work, to explore what kind of worker behavior is related to the learning effect, which can be an interesting problem to human factor [25]. We also shall investigate with other datasets and

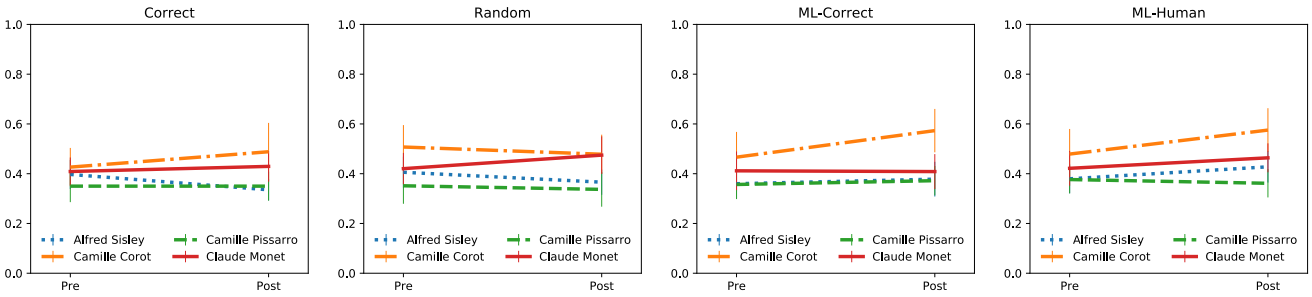

図 7 Learning effect in each class: Quality improvement between pre-test and post-test. Y-axis indicates accuracy rate. Accuracy rate increased particularly for the “Corrot” class in both “ML-Correct” and “ML-Human”, while those did not much increase in “Correct” in case where even perfect accuracy reference were presented.

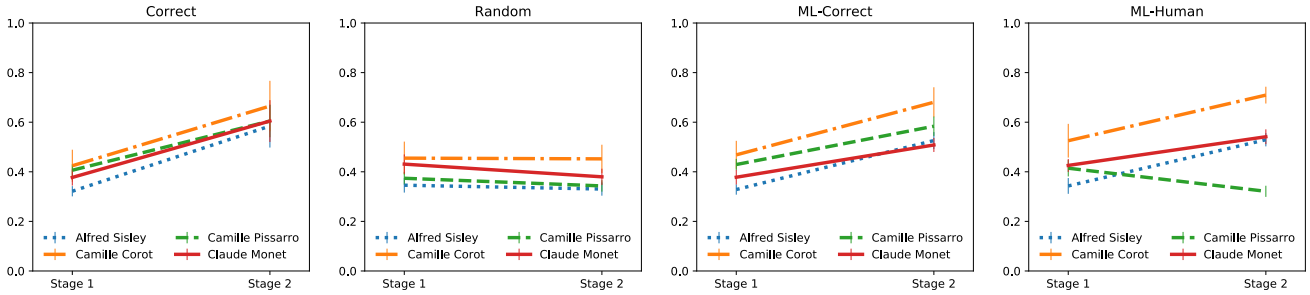

図 8 Self-correction effect in each class: Quality improvement between stage 1 and stage 2. Y-axis indicates accuracy rate. Presenting higher accuracy reference answers gives larger difference. Since the prediction accuracy was 21.7% with “Pissarro” in “ML-Human”, the accuracy rate decreased in that group.

will consider the interpretability of machine prediction [26]?[30] to utilize the learning effects.

Acknowledgments

This work was partially supported by JST CREST GrantNumber JPMJCR16E3 including AIP challenge program, Japan.

文　　献

[1] Kobayashi Masaki. Morishima Atsuyuki. Matsubara, Masaki. A learning effect by presenting machine prediction as a reference answer in self-correction. In *Proceedings of the Second IEEE Workshop on Human-in-the-loop Methods and Human Machine Collaboration in BigData (IEEE HMData2018)*, pp. 3521–3527, 2018.

[2] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, Vol. 51, No. 1, p. 7, 2018.

[3] Adish Singla, Ilija Bogunovic, Gábor Bartók, Amin Karbasi, and Andreas Krause. Near-optimally teaching the crowd to classify. In *ICML*, pp. 154–162, 2014.

[4] Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. Toward a learning science for complex crowdsourcing tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 2623–2634. ACM, 2016.

[5] Steven Dow, Anand Kulkarni, Scott Klemmer, and Bj¨orn Hartmann. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW ’12, pp. 1013–1022, New York, NY, USA, 2012. ACM.

[6] Aniket Kittur. Crowdsourcing, collaboration and creativity. *XRDS: crossroads, the ACM magazine for students*, Vol. 17, No. 2, pp. 22–26, 2010.

[7] Nihar Shah and Dengyong Zhou. No oops, you won’t do it again: Mechanisms for self-correction in crowdsourcing. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, Vol. 48 of *Proceedings of Machine Learning Research*, pp. 1–10, New York, New York, USA, 20–22 Jun 2016. PMLR.

[8] Masaki Kobayashi, Hiromi Morita, Masaki Matsubara, Nobuyuki Shimizu, and Atsuyuki Morishima. An empirical study on short-and long-term effects of self-correction in crowdsourced microtasks. In *HCOMP*, pp. 79–87, 2018.

[9] Mohammad Allahbakhsh, Boualem Benatallah, Aleksandar Ignjatovic, Hamid Reza Motahari-Nezhad, Elisa Bertino, and Schahram Dustdar. Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing*, Vol. 17, No. 2, pp. 76–81, 2013.

[10] Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 614–622. ACM, 2008.

[11] Hongwei Li, Bo Zhao, and Ariel Fuxman. The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing. In *Proceedings of the 23rd international conference on World wide web*, pp. 165–176. ACM, 2014.

[12] Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. *ICWSM*, Vol. 11, pp. 17–21, 2011.

[13] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. Revolt: Col-

laborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI 2017)*. ACM Association for Computing Machinery, May 2017.

[14] Ryan Drapeau, Lydia B Chilton, Jonathan Bragg, and Daniel S Weld. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*, 2016.

[15] Ryo Suzuki, Niloufar Salehi, Michelle S. Lam, Juan C. Marroquin, and Michael S. Bernstein. Atelier: Repurposing expert crowdsourcing tasks as micro-internships. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pp. 2645–2656, New York, NY, USA, 2016. ACM.

[16] Xiaojin Zhu. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *AAAI*, pp. 4083–4087, 2015.

[17] Edward Johns, Oisin Mac Aodha, and Gabriel J Brostow. Becoming the expert-interactive multi-class machine teaching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2616–2624, 2015.

[18] Azad Abad, Moin Nabi, and Alessandro Moschitti. Autonomous crowdsourcing through human-machine collaborative learning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pp. 873–876, New York, NY, USA, 2017. ACM.

[19] Oisin Mac Aodha, Shihan Su, Yuxin Chen, Pietro Perona, and Yisong Yue. Teaching categories to human learners with visual explanations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3820–3828, 2018.

[20] Lev Vygotsky. Zone of proximal development. *Mind in society: The development of higher psychological processes*, Vol. 5291, p. 157, 1987.

[21] Allan Collins, John Seely Brown, and Susan E Newman. Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. *Knowing, learning, and instruction: Essays in honor of Robert Glaser*, Vol. 18, pp. 32–42, 1989.

[22] Jiyi Li, Yukino Baba, and Hisashi Kashima. Hyper questions: Unsupervised targeting of a few experts in crowdsourcing. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 1069–1078. ACM, 2017.

[23] Jerry Alan Fails and Dan R Olsen Jr. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pp. 39–45. ACM, 2003.

[24] Gabriel Viglensoni, Jorge Calvo-Zaragoza, and Ichiro Fujinaga. An environment for machine pedagogy: Learning how to teach computers to read music. In *Proceedings of the IUI Workshop on Music Interfaces for Listening and Creation, Tokyo, Japan*, pp. 7–11, 2018.

[25] Sihem Amer-Yahia and Senjuti Basu Roy. Human factors in crowdsourcing. *Proceedings of the VLDB Endowment*, Vol. 9, No. 13, pp. 1615–1618, 2016.

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144. ACM, 2016.

[27] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929, 2016.

[28] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[29] Hirotaka Akita, Kosuke Nakago, Tomoki Komatsu, Yohei Sugawara, Shin-ichi Maeda, Yukino Baba, and Hisashi Kashima. Bayesgrad: Explaining predictions of graph convolutional networks. *arXiv preprint arXiv:1807.01985*, 2018.

[30] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.