S & M 2358

Using Machine Learning to Estimate Difficulty Levels of Problems

Makoto Koshino[*] and Takuya Koizumi

National Institute of Technology, Ishikawa College, Kitacyujo, Tsubata, Ishikawa 929-0392, Japan

(Received January 22, 2020; accepted May 12, 2020)

Keywords: subjective difficulty, wearable device, machine learning, head movement

In an e-learning environment in which a teacher cannot interact directly with a student, it can be difficult to ascertain a student's difficulty with a subject. In this study, machine learning was used to estimate the level of difficulty of problems experienced by a student to ensure that problems of appropriate difficulty are provided. JINS MEME smart eyewear was used to measure the head movements of students and their results were used to estimate the subjective difficulty that they experienced. Our experimental tests demonstrate the F_1-scores of machine learning for 10 users who were given calculation, kanji (Chinese characters), and programming problems. The feature importance scores of the random forest (RF) were calculated, and the dependence of F_1-score on the type of user was examined. It was found that the mean of the yaw angle was the most important feature in all cases, indicating that the horizontal rotation of the head may depend on the difficulty of the problem.

1. Introduction

In order for classes to progress in a flexible manner, teachers in classrooms should be able to easily understand the subjective level of difficulty a student has with a certain topic. Generally, in one-on-one classes, teachers can adjust the level of difficulty on the basis of a student's facial expressions and gestures, allowing classes to proceed according to the student's ability. However, learning environments such as e-learning and remote classrooms can make it relatively difficult for instructors to accurately determine difficulty levels.

Learning through web-based teaching materials such as e-learning videos has been introduced in several educational fields in recent years. However, analyzing the situation of students is a difficult task for the teachers. For this reason, e-learning requires the ability of the student to effectively understand the context. Ohkawauchi *et al.*[1] investigated the estimation of subjective difficulty experienced by students in the case of lecture videos for e-learning and demonstrated that actions such as pausing and rewinding are correlated with subjective difficulty. Nakamura *et al.*[2] studied the estimation of the subjective difficulty of students during e-learning using a camera to obtain facial characteristics (such as face tilt angles, gaze positions, and whisper time) from images of students' faces. The level of subjective

[*]Corresponding author: e-mail: koshino@ishikawa-nct.ac.jp

https://doi.org/10.18494/SAM.2020.2774

difficulty was estimated using a support vector machine (SVM). Studies on the estimation of student behaviors and the state of web-based lectures and e-learning resources have also been conducted.(3–15) These studies estimated the level of subjective difficulty by measuring eye movements, which were less affected than other phenomena by differences between individuals.(3,4) Shigeta *et al.*(3) studied the level of subjective difficulty of English listening by analyzing data collected by estimating eye movements. The data were gathered using the Freeview software developed by Takei Kikai Kogyo Co., Ltd. The results indicated significant differences in eye movement speed, gaze time, and number of blinks among learners. Okoso *et al.*(4) studied the subjective difficulty of English words in English documents using a deep learning approach in combination with gaze information measured using a Tobii eye tracker. Hence, it is possible to provide a word-based exam appropriate for the level of understanding of an individual. However, in both these studies,(3,4) a dedicated device was required to measure eye movements, which can only be measured when the person is seated in front of a personal computer. Therefore, in the present paper, we consider the use of a wearable device that is not limited to use only in front of a computer, removing the necessity for a camera. The purpose of this study was to use machine learning techniques to measure the head movements of students using a glasses-type wearable device to estimate the subjective difficulty they experienced.

2. Materials and Methods

2.1 JINS MEME

JINS MEME smart eyewear was used in this study as it has nearly the same design and feel as ordinary glasses. There are three types of JINS MEME eyewear: (1) MT, which can measure acceleration and angular velocity; (2) ES_R, which can measure electrooculogram raw data in addition to acceleration and angular velocity; and (3) ES, in which the installed sensor is the same as that of ES_R but no raw electrooculogram data can be collected. ES can measure the speed and strength of blinks using the JINS MEME application programming interface (API).

Studies using JINS MEME have been conducted.(16–18) Ogawa *et al.*(16) studied the estimation of workload by measuring blink data and utterance data when practicing a video game (Tetris) by varying the workload. Nagao *et al.*(17) studied the various states of students when they are learning, such as listening or note taking. In a study of subjective difficulty estimation using JINS MEME, Mori *et al.*(18) considered four-choice questions on English vocabulary: in their study, to construct a system that supports efficient self-study, JINS MEME and the chest-mounted device myBeat, which can measure characteristics such as heart rate and RR interval (RRI), were used for subjective difficulty estimation; the time required for answering was added to the information obtained from JINS MEME ES_R and myBeat, and estimation was performed on the basis of these features.

In the present study, ES was adopted to verify the possibility of estimating difficulty from converted data from the JINS MEME API instead of raw data. Each sensor value of the ES can be recorded at a sampling frequency of 20 Hz using an application connected to an iOS or an Android device via Bluetooth.

2.2 Machine learning

The features used to estimate degree of difficulty have 30 dimensions (six features multiplied by five basic statistics). The six features consist of head movements: means of *x*-, *y*-, and *z*-axis acceleration, roll, pitch, and yaw angle over a time window of 2 s. The five basic statistics consist of mean, standard deviation, maximum value, minimum value, and median. A time window of 2 s is often used in studies employing acceleration sensors. The resulting 30-dimensional feature is standardized to have a mean of zero and a variance of one, and machine learning is performed using a 10-dimensional feature through principal component analysis (PCA). Four methods were used: SVM, random forest (RF), decision tree (DT), and k-nearest neighbor (k-NN). DT and RF were used because their classification rules are easy to understand and the corresponding data visualization is also convenient, and SVM can perform binary classification with high accuracy, as has been reported in previous studies.(2,3,18) k-NN, which is the most basic classification method, was used owing to its simplicity.

The JINS MEME ES enables measurement of the speed and strength of a blink as well as the acceleration and angular velocity of the head. First, we measured the blinks and head movement and used these to perform estimations using machine learning. However, the number of blinks is known to generally decrease from 20 times per minute in normal activity to 10 times a minute during reading and five times a minute when working on personal computers. Therefore, blinks were often not detected when the time window was 2 s. Thus, the time window was set to 20 s, and the feature importance score of RF was calculated. It was found that the feature of the head movement was the most important and that the number of blinks did not provide any useful insights. For this reason, we excluded blink data from the features described in this study.

2.3 Experimental methods

Studies focused on calculation problems(2) and English language problems(3,4,18) that estimated learners' difficulty levels have been conducted, and JINS MEME eyewear has also been utilized.(18) In this study, we considered three types of problems, namely, calculation problems and kanji (Chinese characters) problems, whose difficulty levels can be easily adjusted, and programming problems, which are likely to generate individual differences with regard to ability. The calculation problems were fill-in-the-missing-number problems. An example of an easy calculation problem is "■ ÷ 3 = 1" and that of a difficult problem is "(■ − 9 / 6) = 3 ÷ 0.125". The kanji problems consisted of writing the corresponding kanji of words given in hiragana. For the kanji problems, we used the third and seventh levels of the Japan Kanji Aptitude Test. The programming problem involved the written part of a university entrance examination. Figure 1 shows a view of the experiment. We recruited 10 participants, all of which were males aged between 20 and 22 years. All participants were informed of the purpose and content of the study and agreed to privacy protection. Furthermore, the Research Ethics Board of the National Institute of Technology, Ishikawa College approved this study through an ethics review.

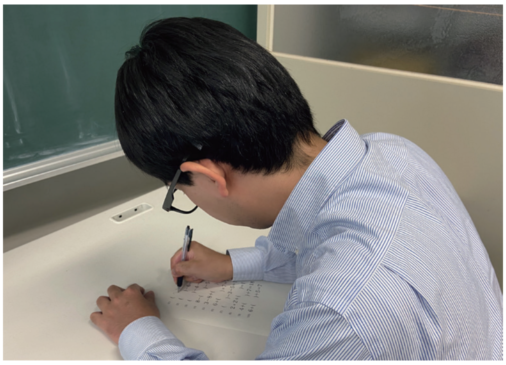

Fig. 1. (Color online) A view of the experiment.

The experimental procedure for the calculation problems was as follows:

(1) Each participant was made to sit on a chair and wear the JINS MEME eyewear.

(2) They were asked to solve a problem within a time of 6 min.

(3) They were then asked to take a break for 3 min.

(4) They were then asked to solve another problem within 6 min.

(5) After completion of the tasks, the participants were asked to state whether step (2) or step (4) was more difficult. The choices were labeled as easy or difficult.

These steps were repeated for the kanji and programming problems.

The first minute of the 6-min-long test data that were recorded was excluded because the corresponding data were unstable. We prepared several problems that could not be solved in 6 min, and none of the participants were able to solve them. The data from the second to the fifth minute were categorized as training data (number of data: 240), and the data from the last minute were categorized as test data (number of data: 60). When data are randomly sampled and trained, there is the potential for data leakage; we selected our experimental method to avoid this problem. The training data were examined to find the parameter that maximizes the score of 10-fold cross-validation (CV), and the model was trained using this parameter. Next, the F_1-score (also F-score or F-measure), which is a measure of a test's accuracy, was verified to evaluate the estimations.

3. Results

3.1 F_1-score for each user

The experimental data were used to evaluate the estimation for each user. The results obtained are presented in Tables 1–3. The rows in Tables 1–3 represent the users and the columns represent the learning methods.

For the calculation problem, the F_1-scores were 85% with SVM, 89% with RF, 80% with DT, and 81% with k-NN. However, in the case of user J, the average F_1-score was 45%, which was lower than that of the other users. In the case of the kanji problems, the F_1-scores were 90% with SVM, 87% with RF, 83% with DT, and 88% with k-NN. In the programming problems,

Table 1
F_1-score for each user (calculation problems).

	SVM	RF	DT	k-NN	Avg.
A	98%	100%	78%	80%	89%
B	83%	100%	100%	77%	90%
C	92%	88%	82%	92%	88%
D	95%	97%	55%	73%	80%
E	62%	77%	85%	68%	73%
F	100%	100%	100%	97%	99%
G	93%	98%	98%	87%	94%
H	82%	92%	58%	82%	78%
I	100%	98%	100%	100%	99%
J	47%	37%	42%	57%	45%
Avg.	85%	89%	80%	81%	

Table 2
F_1-score for each user (kanji problems).

	SVM	RF	DT	k-NN	Avg.
A	100%	92%	95%	95%	95%
B	55%	70%	67%	78%	67%
C	65%	62%	63%	58%	62%
D	92%	93%	88%	87%	90%
E	100%	92%	82%	98%	93%
F	98%	97%	92%	100%	97%
G	100%	95%	85%	93%	93%
H	92%	77%	68%	80%	79%
I	97%	98%	90%	95%	95%
J	100%	97%	97%	100%	98%
Avg.	90%	87%	83%	88%	

Table 3
F_1-score for each user (programming problems).

	SVM	RF	DT	k-NN	Avg.
A	92%	72%	77%	87%	82%
B	73%	77%	73%	67%	73%
C	62%	55%	48%	68%	58%
D	77%	65%	83%	65%	73%
E	100%	97%	97%	93%	97%
F	80%	97%	58%	77%	78%
G	60%	60%	60%	65%	61%
H	80%	68%	65%	78%	73%
I	92%	100%	97%	93%	95%
J	52%	63%	62%	48%	56%
Avg.	77%	75%	72%	74%	

Table 4
Important features (calculation problems).

Feature	No. of users
Mean of yaw angle	9 people
Minimum value of yaw angle	8 people
Median of yaw angle	5 people
Maximum value of yaw angle	4 people
Maximum value of *y*-axis acceleration	2 people
Mean of roll angle	1 person
Minimum value of *x*-axis acceleration	1 person

Table 5
Important features (kanji problems).

Feature	No. of users
Mean of yaw angle	10 people
Minimum value of yaw angle	6 people
Median of yaw angle	5 people
Maximum value of yaw angle	5 people
Maximum value of *z*-axis acceleration	2 people
Mean of roll angle	1 person
Minimum value of *x*-axis acceleration	1 person

Table 6
Important features (programming problems).

Feature	No. of users
Mean of yaw angle	8 people
Minimum value of yaw angle	8 people
Median of yaw angle	7 people
Maximum value of yaw angle	1 people
Mean of roll angle	1 people
Minimum of roll angle	1 person
Maximum value of pitch angle	1 person
Minimum value of *y*-axis acceleration	1 person
Mean of *z*-axis acceleration	1 person
Maximum value of *z*-axis acceleration	1 person

the F_1-scores were 77% with SVM, 75% with RF, 72% with DT, and 74% with k-NN, and for several users, the F_1-score was more than 70%. However, the F_1-score was low for users C, G, and J.

Feature importance of RF was evaluated, and consequently, the three most important features were recorded for each user, with the results presented in Tables 4–6. It was found that the mean of the yaw angle is the most important feature in all cases. This indicates that the horizontal rotation of the head may depend on the difficulty of the problem. It appears that this was because the speed of problem-solving depends on the level of difficulty of the problem. In general, when solving a problem, the head tilts and nods, and therefore, the pitch angle, which represents the vertical movement of the head, is likely to be related to the level of difficulty. However, in this experiment, only a few users exhibited this tendency, and therefore, the yaw angle was found to be the most important feature.

3.2 Dependence of F_1-score on the type of user

The dependence of F_1-score on the type of user was evaluated using experimental data. In particular, the model was trained using data from each user (number of data: 300), and F_1-scores for all users' test data were evaluated accordingly. The SVM is only used for evaluation, and the F_1-scores are presented in Tables 7–9. It was observed that the responses to calculation and programming problems vary greatly among individuals. Depending on the user, the F_1-score was as high as 100% or as low as 1%. However, the kanji problems appeared to be easiest to characterize using the users' results. For users A–H, the F_1-score was at least 50%, and in most

Table 7
Dependence of F_1-score on the type of user (calculation problems).

		Users (test data)									
		A	B	C	D	E	F	G	H	I	J
Users (training data)	A	—	98%	43%	5%	18%	15%	81%	23%	97%	40%
	B	100%	—	29%	4%	12%	1%	73%	17%	90%	35%
	C	52%	36%	—	44%	81%	63%	76%	73%	52%	66%
	D	1%	3%	64%	—	90%	97%	14%	76%	20%	69%
	E	13%	3%	72%	98%	—	100%	57%	85%	18%	70%
	F	20%	8%	74%	95%	92%	—	41%	76%	31%	72%
	G	85%	83%	58%	7%	35%	11%	—	44%	62%	44%
	H	44%	50%	55%	36%	51%	60%	47%	—	46%	58%
	I	95%	76%	38%	48%	43%	71%	46%	42%	—	47%
	J	43%	60%	81%	45%	74%	53%	77%	49%	41%	—

Table 8
Dependence of F_1-score on the type of user (kanji problems).

		Users (test data)									
		A	B	C	D	E	F	G	H	I	J
Users (training data)	A	—	92%	74%	94%	97%	100%	94%	35%	18%	0%
	B	99%	—	79%	91%	97%	100%	91%	43%	26%	3%
	C	91%	74%	—	94%	97%	98%	96%	36%	39%	2%
	D	97%	80%	80%	—	100%	98%	100%	55%	25%	7%
	E	100%	83%	82%	99%	—	100%	99%	45%	18%	0%
	F	100%	88%	81%	96%	100%	—	99%	44%	22%	0%
	G	95%	81%	81%	96%	99%	97%	—	49%	22%	2%
	H	96%	88%	59%	78%	75%	96%	72%	—	17%	11%
	I	20%	25%	40%	28%	17%	8%	41%	23%	—	75%
	J	0%	13%	21%	4%	0%	0%	0%	63%	78%	—

Table 9
Dependence of F_1-score on the type of user (programming problems).

		Users (test data)									
		A	B	C	D	E	F	G	H	I	J
Users (training data)	A	—	46%	49%	43%	31%	53%	62%	64%	26%	57%
	B	36%	—	40%	23%	31%	74%	21%	74%	87%	48%
	C	49%	45%	—	90%	100%	5%	86%	38%	2%	58%
	D	55%	19%	83%	—	97%	14%	88%	36%	3%	56%
	E	47%	38%	91%	86%	—	7%	90%	37%	2%	57%
	F	30%	82%	14%	10%	0%	—	9%	60%	100%	29%
	G	51%	25%	58%	84%	89%	17%	—	31%	7%	62%
	H	44%	56%	45%	53%	12%	75%	24%	—	53%	39%
	I	48%	80%	14%	1%	0%	93%	9%	63%	—	33%
	J	75%	15%	88%	93%	99%	1%	90%	35%	1%	—

cases, the F_1-score was at least 80%. Estimations for users A–H were less accurate than those for users I and J. In addition, the results depended on the type of problem. In a kanji problem, participants are unable to calculate the solution; they either know or do not know the answer. We suppose that the movement of these users was similar because it depends on the problem's difficulty. In such a case, the F_1-score of difficulty estimation is more dependent on the type of user. Therefore, we conclude that user dependence may be reduced by employing learning data from users with similar abilities.

4. Conclusions

In this study, we used a machine learning approach to estimate the degree of difficulty experienced with different types of learning content based on the head movements of students. We used JINS MEME eyewear, which does not require a camera, to track the head movements of students and estimate the difficulty of problems. Our results show the high F_1-score of the proposed approach. The most important features of RF were examined, and the yaw angle, which represents the left–right head rotation, was found to be the most important feature in all the cases. Additionally, when the dependence of the F_1-score on the type of user was examined using other training models, we observed significant differences in the results depending on the particular student and type of problem. For future work, the F_1-score will be examined by increasing the number of participants and considering other factors such as age, gender, and ability. This will enable us to develop a highly accurate approach to determining difficulty levels using machine learning, which may be applied in the rapidly expanding field of online learning.

Acknowledgments

This work was supported by JSPS KAKENHI (grant number: JP16K01097).

References

1 T. Ohkawauchi, J. Ohya, S. Yonemura, and Y. Tokunaga: Jpn. Soc. Edu. Technol. **36** (2012) 193 (in Japanese).

2 K. Nakamura, K. Kakusho, and M. Murakami: IEICE Trans. Inf. Syst. (Japanese edition) **93** (2010) 568 (in Japanese).

3 A. Shigeta, K. Hamamoto, and K. Nosu: Trans. Inst. Electr. Eng. Jpn. C **131** (2011) 800 (in Japanese).

4 A. Okoso, S. Ishimaru, O. Augereau, and K. Kise: Tech. Rep. IEICE **116** (2017) 187 (in Japanese).

5 Z. Zhu, S. Ober, and R. Jafari: Proc. IEEE 14th Int. Conf. Wearable and Implantable Body Sensor Networks (BSN) (IEEE, 2017) 13–18. https://doi.org/10.1109/BSN.2017.7935996

6 T. Yamane, K. Nakamura, M. Ueda, M. Mukunoki, and M. Minoh: SIG Adv. Learning Sci. Technol. **60** (2010) 7 (in Japanese).

7 F. Murai K. Kakusho, T. Kojima, and M. Murakami: Jpn. Soc. Inf. Syst. Edu. **32** (2015) 48 (in Japanese).

8 D. Shimada and H. Iyatomi: Jpn. Soc. Fuzzy Theory Intell. Inf. **29** (2017) 517 (in Japanese).

9 M. Sakaguchi, M. Toyoura, X. Mao, and M. Hanawa: IPSJ SIG Tech. Rep. CVIM 2017-CVIM-**209** (2017) 1 (in Japanese).

10 M. Sakaguchi, M. Toyoura, D. Akoh, X. Mao, S. Nishiguchi, M. Hanawa, and M. Murakami: Proc. Annu. Conf. Japanese Society for Information and Systems in Education, C4–1 (2017) (in Japanese).

11 X. Zhang, C. W. Wu, P. Fournier-Viger, L. D. Van, and Y. C. Tseng: Proc. IEEE 18th Int. Symp. A World of Wireless, Mobile and Multimedia Networks (WoWMoM) (IEEE, 2017) 1–9.

12 T. Tezuka, Y. Seino, R. Furutani, and T. Satoh: Jpn. Soc. Fuzzy Theory Intell. Inf. **28** (2016) 952 (in Japanese).

13 S. Matsumoto, R. Wataya, D. Iwai, and K. Sato: IEEJ Trans. Electron. Inf. Syst. **133** (2012) 1 (in Japanese).

14 K. Li, T. Kumazaki, and M. Saigusa: Jpn. Soc. Inf. Syst. Edu. **33** (2016) 110 (in Japanese).

15 Y. Asai, D. Aikawa, and H. Egi: IPSJ Interaction **2B-35** (2019) 595 (in Japanese).

16 T. Ogawa, M. Takahashi, and R. Kawashima: IFAC-PapersOnLine **49** (2016) 331. https://doi.org/10.1016/j.ifacol.2016.10.571

17 M. Nagao, H. Saito, F. Taniguchi, and Y. Sase: Memoirs of Hokkaido Information University **29** (2017) 47 (in Japanese).

18 T. Mori and T. Hasegawa: Proc. IEEE TENCON **2018** (IEEE, 2018) 1045–1050. https://doi.org/10.1109/TENCON.2018.8650315