LETTER

Investigation on e-Learning Status Estimation for New Learners—Classifier Selection on Representative Sample Selection

Siyang YU[†a)], ***Nonmember***, **Kazuaki KONDO**[†], **Yuichi NAKAMURA**[†], ***Members***, **Takayuki NAKAJIMA**[††], ***and*** **Masatake DANTSUJI**[†], ***Nonmembers***

SUMMARY This article introduces our investigation on learning state estimation in e-learning on the condition that visual observation and recording of a learner's behaviors is possible. In this research, we examined methods of adaptation for a new learner for whom a small number of ground truth data can be obtained.

key words: *e-Learning status estimation, visual sensing, interpersonal differences, adaptation to new learner*

1. Introduction

If learning is adaptive and customized to meet the needs of each learner, learning can be effective and efficient. However, the lack of feedback regarding how learning occurs in e-learning constrains the personalization and adaptation. To effectively address this issue, there is a tremendous demand for measuring the learners' state in e-learning. Many previous studies have focused on the recognition of learning states [1]–[5]. Different affective-cognitive states and sensing modalities have been investigated based on different perspectives. Various features have been proved to be useful clues for observing and evaluating a learner's learning state. However, the features are heavily dependent on personal characteristics, and evaluating the learning state of an individual learner is substantially affected by the interpersonal differences, especially when characteristics are not known in advance. This issue was experimentally confirmed in [5], which proposed a method for selecting suitable classifiers to estimate the state of learning based on the set of pre-trained classifiers. Although a degree of improvement in learning was observed, there are some areas for further improvements. This study aims to investigate a method for choosing an appropriate classifier for estimating learning states, which addresses the problem of choosing a small number of samples and selecting a suitable classifier.

Manuscript received March 5, 2019.
Manuscript revised October 2, 2019.
Manuscript publicized January 20, 2020.
[†]The authors are with Academic Center for Computing and Media Studies, Kyoto University, Kyoto-shi, 606–8501 Japan.
[††]The author is with Graduate School of Human and Environmental Studies, Kyoto University, Kyoto-shi, 606–8501 Japan.
a) E-mail: yusiyang@ccm.media.kyoto-u.ac.jp
DOI: 10.1587/transinf.2019EDL8043

2. Objectives and Problem

2.1 Background Conditions

We assume that the same process is performed for obtaining the learning state classifiers in reference [5]. This process comprises the following steps.

(a) We collected behavior samples during e-learning from learners. In particular, nonverbal behaviors were recorded through visual sensing without imposing any type of constraint or intrusion on the learners. Examples of captured images are depicted in Fig. 1.

(b) The recorded video data were divided into 30-second segments. Facial recognition is employed for each segment, and head movements, facial features, and gazing features are automatically detected. These detected features comprise a 33-dimensional feature vector, as demonstrated in Fig. 2.

(c) A small number of collaborators called prototype learners annotate the ground truth of learning states for their segments by introspection. The learning states are indexes on concentration-distraction, difficulty-ease, and interest-boredom, which were scored on a five-point scale.

(d) A support vector machine (SVM) for estimating the scores was trained by applying the feature vectors obtained in (b) and the ground truth of learning states in (c). The learning states in each segment can be estimated using the trained SVMs.

In reference [5], this process was conducted for 3119 samples by seven collaborators. Based on the condition that training data and testing data came from the same learner, an accuracy of 60.9%-65.6% for exact matching and 86.2%-92.7% for lenient matching was obtained.

2.2 Study Objectives

This study aims to obtain a good classifier for a new learner,

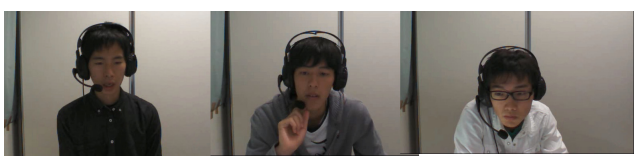

Fig. 1 Examples of captured images

Feature source:	Feature items:
Presence information	Present proportion
	Distance (Max, Average, Min)
Head and facial parts information	Face detection successful proportion
	Lips movement (Max, Average, Min)
	Eyebrows movement (Max, Average, Min)
	Head pose angle Pitch (Max, Average, Min)
	Head pose angle Yaw (Max, Average, Min)
	Head pose angle Roll (Max, Average, Min)
	Head position X-coordinate (Max, Average, Min)
	Head position Y-coordinate (Max, Average, Min)
	Head position Z-coordinate (Max, Average, Min)
Probability of gazing at screen	High probability proportion
	Moderate probability proportion
	Low probability proportion
	Consider as zero probability proportion

Fig. 2 33-dimensional feature vector

i.e., a non-prototype learner, given the following assumptions. Sufficient data, i.e., records of learning behaviors and ground truth of learning state, are obtained through the collaboration of the prototype learners. Additionally, the learning behaviors of a new learner are automatically recorded. Meanwhile, the collection of numerous ground truth enough to train SVM is not achievable due to the required effort and time. Only very few "representative samples" can be annotated with learning states by each new learner. To estimate a new learner's learning states, we alternatively utilize classifiers that are pre-trained by employing the dataset of prototype learners. Various classifiers can be obtained by including or excluding certain samples from the prototype learners.

This idea presents a new problem of choosing an appropriate classifier that demonstrates good performance for a new learner. As one method to solve this problem, in reference [5], five representative samples were randomly chosen from a new learner's data, and a classifier that demonstrated the best performance for the representative samples was chosen. By harnessing this method, the selected classifier mostly demonstrated inferior performance when compared with the classifier with the best performance out of all the pre-trained classifiers. Therefore, the reduction of the gap between the selected classifier and that with the best performance for a new learner is an important objective.

2.3 Problem Statement

The following demonstrates a concise description of the problem. (i) A sufficient number of samples with ground truth from the prototype learners are given. (ii) An automatic observation of a new learner is possible in the same way that the prototype learners are observed. (iii) The learning state scores can be given only to few representative samples for each new learner.

The challenge is to determine how to choose representative samples that result in better performance. The method for selecting a classifier was accuracy-based method because it demonstrated better performance compared to the other methods employed in [5]. It simply chooses the classifier(s) that demonstrate(s) the best accuracy for a set of representative samples.

3. Scheme for Selecting Representative Samples

3.1 Basic Assumptions

The proposed method of choosing representative samples is based on the following assumptions.

A1: Frequently appearing samples can be good representative samples. Given a representative sample that is correctly classified, neighboring samples of a representative sample are also expected to be correctly classified if the self-scoring is consistent throughout the learning period. A representative sample that lies in a region of higher occurrence probability has more neighboring samples, which are expected to be correctly classified.

A2: A set of representative samples that covers a wider area of the feature space provides better accuracy. Similar to A1, not only samples neighboring to representative samples, but also in-between representative samples have a better probability of being correctly classified compared to samples far from representative samples.

A3: A set of representative samples with enough variety of classes gives better accuracy. Samples with different scores may correspond to different behaviors. It is important to obtain enough variation in the classes of a set of representative samples.

3.2 Representative Sample Selection Based on Assumptions

According to the above assumptions, a set of samples with high occurrence probability are given priority as representative samples, and the following scheme is proposed: Step 1: Estimate the probability distribution of samples obtained from a new learner. Step 2: Apply clustering to samples with a large probability density. Step 3: Collect representative samples by choosing one sample from each cluster. Step 4: Select the classifier that demonstrates the best performance for the set of selected samples.

For Step 1, kernel density estimation with Gaussian kernel [6] is applied to the samples of a new learner. In this process, the samples in which no face was detected are excluded because they are less important in estimating the learning state.

$$\rho f(x) = \sum_{i=1}^{N} K(\frac{x - x_i}{h})$$

$$K(x; h) \propto exp(-\frac{x^2}{2h^2})$$

Where ρ is a coefficient, K represents the Gaussian kernel function, x_i is the observation $(x_1, x_2, \ldots, x_N)$, and h is the smoothing parameter referred to as the bandwidth.

For Step 2, hierarchical agglomerative clustering [6] is applied to samples with a high probability density. Clustering is not applied to all the samples at once because it is difficult to know the appropriate number of clusters for a large number of samples. Clustering is also avoided if the results would be poor with respect to the growth of the sample population. In Step 3, the sample corresponding to the highest probability density is chosen as the representative sample from each cluster. After the selection, each new learner is asked to provide ground truth scores of learning states for the selected representative samples. Then, in Step 4, the accuracy-based method is applied to these samples to select the suitable classifier.

4. Experimental Results and Discussion

Experiments were conducted to verify assumptions and the scheme for classifier selection, and as a test bed, the dataset gathered in the previous research [5] was used. In the dataset, 3119 samples from seven participants were collected, with each data point being an observed feature vector and the ground truth score of learning states. A total of 63 classifiers were trained, one of which was the unified classifier that is trained by all the samples, while the others are trained by one participant's data or by two or more participants' data.

Leave-one-out cross validation was applied for verification (i.e., each participant was considered to be a new learner), and the others were considered as prototype learners. Steps 1 through 4 were applied to the data of each new learner to obtain five representative samples, and a suitable classifier was selected for each of the three learning states. The selected classifier was applied to all the samples of the new learner, and the accuracy was calculated for both strict and lenient matching conditions.

Samples with the top 10-40% of high occurrence probability were chosen as clustering targets. An example of dendrogram in hierarchical clustering for a new learner is shown in Fig. 3, which illustrates how the samples are hierarchically clustered from the bottom to the top. The horizontal axis indicates the sample index, and the vertical axis represents the distance. Five clusters were identified within this type of hierarchical structure at the height of the dotted line.

The average accuracy of the selected classifiers with five representative samples (RS) is shown in Fig. 4. The average accuracy of the selected classifiers with ten RS is shown in Fig. 5.

The performance obtained by randomly choosing RS is presented as the baseline for comparison, which is the average of 2000 random selection trials per learner. The range between max and min value of new learners is shown too. The best overall performance among the 63 classifiers is also presented, which is indicative of the upper bound of performance.

The improvements of the average accuracy are not significant with the proposed scheme, while the performance is slightly better than that of the random selection method. Meanwhile, the proposed method has a good characteristic for educational service. It deterministically obtains the pre-

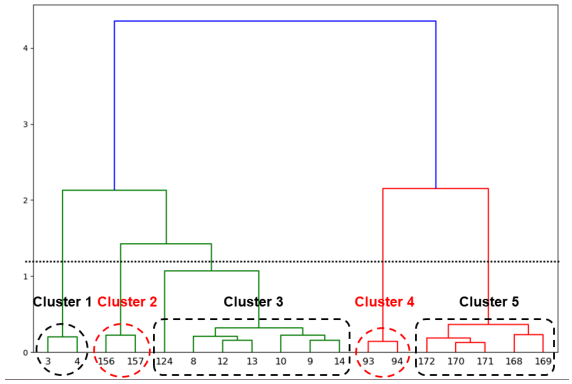

Fig. 3 Example of dendrogram in hierarchical clustering

5 Representative Samples		Unified Classifier	Random Selection			Clustering Selection	Best Classifier
			Max over 7 learners	Average for 7 learners	Min over 7 learners (worst case)		
Strict Matching	Concentration-Distraction	31.4%	43.4%	35.8%	26.9% (9.3%)	38.6% (24.9%)	48.1%
	Difficulty-Ease	26.4%	34.4%	27.1%	21.4% (8.7%)	27.5% (20.1%)	35.9%
	Interest-Boredom	41.7%	57.7%	40.9%	22.5% (7.6%)	41.2% (25.4%)	52.7%
Lenient Matching	Concentration-Distraction	74.3%	86.3%	77.3%	66.9% (32.1%)	77.0% (61.1%)	85.4%
	Difficulty-Ease	69.2%	71.6%	65.8%	59.1% (27.3%)	67.9% (49.2%)	72.6%
	Interest-Boredom	80.2%	91.7%	79.9%	61.8% (36.8%)	78.9% (64.0%)	86.2%

Fig. 4 Classifier selection performance comparison with five RS. The table entries for "Min over 7 learners (worst case)" under the random selection represent the average value of the minimum accuracy over seven learners and the worst accuracy over all the trials. The table entries for "Clustering Selection" also represent the average accuracy for seven learners and the worst accuracy.

10 Representative Samples		Unified Classifier	Random Selection			Clustering Selection	Best Classifier
			Max over 7 learners	Average for 7 learners	Min over 7 learners (worst case)		
Strict Matching	Concentration -Distraction	31.4%	47.0%	38.8%	28.1% (12.6%)	40.1% (24.3%)	48.1%
	Difficulty-Ease	26.4%	39.4%	28.6%	21.9% (9.3%)	26.2% (16.9%)	35.9%
	Interest-Boredom	41.7%	62.8%	43.7%	23.4% (8.4%)	45.1% (26.1%)	52.7%
Lenient Matching	Concentration -Distraction	74.3%	88.4%	79.1%	68.0% (40.0%)	79.0% (61.1%)	85.4%
	Difficulty-Ease	69.2%	75.4%	67.2%	59.6% (32.5%)	67.0% (53.7%)	72.6%
	Interest-Boredom	80.2%	93.1%	81.0%	63.0% (35.3%)	81.7% (63.9%)	86.2%

Fig. 5 Classifier selection performance comparison with ten RS. The table entries for "Min over 7 learners (worst case)" under the random selection represent the average value of the minimum accuracy over seven learners and the worst accuracy over all the trials. The table entries for "Clustering Selection" also represent the average accuracy for seven learners and the worst accuracy.

	Accuracy On RS	Accuracy On Clusters	Accuracy On AS
Concentration-Distraction	72.1%	44.5%	38.6%
Difficulty-Ease	51.4%	29.4%	27.5%
Interest-Boredom	81.4%	52.1%	41.2%

Fig. 6 Classification accuracy of different sets of samples.

sented accuracy, which makes it possible to clearly avoid the cases much worse than the average. Figures 4 and 5 depict the accuracy of the worst cases among all the trials by the random selection method. The chances of such bad cases are inevitable in the random selection method. We cannot verify the accuracy of learning state estimation unless the new learners' samples are thoroughly scored by taking the learner's considerable amount of time and efforts. Therefore, the characteristic of the proposed scheme matches with a natural requirement that we should avoid serious disadvantage for any learners even if we lose chances of certain merit for some other learners.

Concerning assumption A1, Fig. 6 shows the accuracy in a 5-cluster case with strict matching conditions for the RS, for the samples within the cluster of each RS, and for the overall samples (AS). It is surprising that the accuracy for the RS is only around 50% regarding the difficulty-ease state, which is much lower than the other two learning states. This implies that behaviors are diverse with respect to the difficulty-ease state of the learner, and a more sophisticated method is needed to deal with them. In addition to this issue, the results indicated that a better accuracy for the samples within the clusters compared to the accuracy for the AS is possible. However, it also implies that the amount of improvement is not enough to achieve significant improvements in the total accuracy.

Concerning assumptions A1, A2, and A3, an improvement in 10-cluster cases compared to 5-cluster cases was observed for concentration-distraction and interest-boredom. With the increased number of RS, more samples of neighboring or in-between RS were obtained. Variations in the RS were also obtained. This effect can be seen based on the improved performance, with the exception of the difficulty-ease state. A possible reason for the worse difficulty-ease performance may be partially due to the diverse behaviors in these cases.

Concerning assumption A2 and A3, a high degree of accuracy is expected if the characteristics of the RS are similar to the characteristics of AS associated with a new learner. To verify this, we examined the relationship between the accuracy of selected classifiers and ground truth score distribution of RS and AS. The distribution of the scores for each learning state can be obtained from ground truth for both the RS and AS, and Pearson correlation was observed between their distributions. The result in strict matching is shown in Fig. 7. Each dot represents a selected classifier. The horizontal axis represents the correlation, and the vertical axis indicates the accuracy of the classifier (i.e., between the worst (0.0) and the best (1.0)). Two or more classifiers are often selected in Step 4, and they are represented by a group of vertically aligned dots.

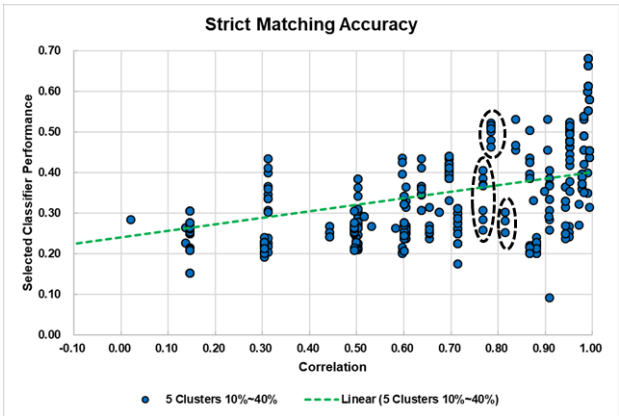

Fig. 7 Tendency of correlation and classifier performance in strict matching. Some of the cases in which multiple classifiers were selected are notified by dashed closures.

A rough relationship between correlation and classifier accuracy was observed, and selected classifiers tended to have a higher degree of accuracy if they had a larger correlation. The linear regression value was 0.16 ($p < 0.01$) for strict matching criterion. It is important to note that the variations among multiple classifiers for each selection were not negligible. With the same correlation value, the accuracy varied among classifiers. If the best one can be chosen for a set of RS, the accuracy would be much higher. The best accuracy increases as the correlation increases. However, an effective method for this purpose was not clearly identified based on the scope and results of this study. Further investigation regarding this issue should be part of future research.

5. Conclusion

In this study, we investigated a scheme for dealing with interpersonal differences in estimating learners' learning states. Specifically, based on our basic assumptions, the selection of RS that commonly appear during e-learning of a new learner was examined. A slight improvement in the average accuracy was observed, although it was not significant. However, the proposed method did demonstrate the advantage of avoiding bad cases. Certain characteristics of the data and classifications were confirmed. Neighboring samples around RS were not classified well, especially in difficulty-ease state. In addition, the distribution of the RS displayed a significant correlation regarding accuracy, which make them a potentially valuable indicator for future investigations of new methods.

References

[1] N.J. Butko, G. Theocharous, M. Philipose, and J.R. Movellan, "Automated facial affect analysis for one-on-one tutoring applications," 2011 IEEE Int. Conf. Automatic Face & Gesture Recognition and Workshops (FG 2011), pp.382–387, IEEE, 2011.

[2] S.K. D'mello, S.D. Craig, A. Witherspoon, B. Mcdaniel, and A. Graesser, "Automatic detection of learner's affect from conversational

cues," User modeling and user-adapted interaction, vol.18, no.1-2, pp.45–80, Feb. 2008.
[3] L. Shen, M. Wang, and R. Shen, "Affective e-learning: Using "emotional" data to improve learning in pervasive learning environment," Educational Technology & Society, vol.12, no.2, pp.176–189, 2009.
[4] B. Woolf, W. Burleson, I. Arroyo, T. Dragon, D. Cooper, and R. Picard, "Affect-aware tutors: recognising and responding to student affect," Int. J. Learning Technology, vol.4, no.3-4, pp.129–164, 2009.
[5] S. Yu, K. Kondo, Y. Nakamura, T. Nakajima, and M. Dantsuji, "Learning state recognition in self-paced e-learning," IEICE Trans. Inf. & Syst., vol.100, no.2, pp.340–349, 2017.
[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, and G. Louppe, "Scikit-learn: Machine learning in python," J. Machine Learning Research, vol.12, no.Oct, pp.2825–2830, 2011.