6ZE-04

Case Study on Evaluation of Students' Discussion Statements' Appropriateness Based on their Heart Rate

Shimeng Peng† Shigeki Ohira‡ Katashi Nagao†
Graduate School of Informatics, Nagoya University†
Information Technology Center, Nagoya University‡

1 Introduction

In this study, we explore how students' Heart Rate (HR) data can be used to evaluate their answer-statements' appropriateness while students completed a Question-and-Answer (Q&A) session in discussions. We adopt Apple Watch to collect students' HR and Web-based scoring method to evaluate answer-statements' appropriateness. HR features were analyzed and used in three machine learning models: logistic regression, support vector machine, and random forest for prediction of answer-statements' appropriateness. Leave-one-student-out cross validation was used to evaluate classifiers' accuracy on all the students. We also take insight into the performance of HR-based prediction models on student groups with different level of experience on the discussion. We validated the effectiveness of our proposed models regarding evaluation of students' discussion statements' appropriateness.

2 Experiments Design and Data Acquisition

2.1 Discussion Experimental Design

We conducted discussion experiments based on our regular seminar-style discussion environment in which a presenter explains a research topic while displaying slides, and Q&A with the meeting participants and presenter is carried out during the presentation. We used Discussion-Mining System (DM) [1] to record all the Q&A segments generated in the experiments for analysis. Fifteen lab members participated in our experiments, including four undergraduates and eight graduate students, and three professors; Each discussion experiment was held once a week for around 1.5-2 hours in which we asked one student to be the presenter and give a presentation on their recent research and the others to raise questions as questioners.

2.2 Data-Acquisition System

In our experiments, we adopted Apple Watch series 2 to collect discussion participants' HR data. Every experiment, we asked one presenter to wear Apple Watch on their left hand before each discussion; The collected HR information is shown on the Apple Watch's screen, as well as synchronously presented on our HR web browser, which we have given more introduction in our previous work [2].

Our starting point was categorizing the answer quality of Q&A segments of discussions into low or high according to how correctly the presenter answered the questioners' questions. We designed a real-time Web-based scoring method shown in Figure 1, which can be viewed on tablets to ask all participants to give timely evaluation scores of not only the answer quality but also the question difficulty after each Q&A segment.

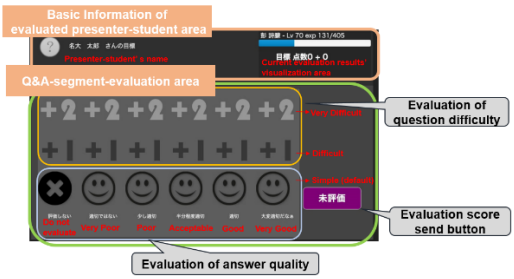

Figure 1: Real-time Web-based scoring page on tablet

There are two main area of this web evaluation page. Basic information of evaluated presenter area, which is used to present the current discussion presenter's name and also to show accumulated evaluated scores he or her obtained. The other area is Q&A-segment-evaluation area: In this area, all participants need to give two evaluation scores after each Q&A segment, one for answer quality with the five-point scale from very poor to very good by tapping different "smile" buttons on the bottom of the page. The answers were considered high quality if the scores were "Good" or "Very Good", and those considered low quality if the scores were the others. Participants had to also evaluate the difficulty of the question of the same Q&A segment by tapping "+1", which means "Difficult", or "+2", which means "Very Difficult", on the top of each smile button. If they tapped nothing, it went to the default evaluation value, i.e., this question was "Simple".

3 Data Analysis

3.1 Experimental Data

Our study was based on a real-lab seminar-style discussion; therefore, our discussion experiments were conducted in the first and second halves of the

心拍数に基づく学生の発言の適切性の評価に関する事例研究
†彭 詩朦 ‡大平 茂輝 †長尾 確
†名古屋大学 大学院情報学研究科
‡名古屋大学 情報基盤センター

academic year: discussion experiments data were collected 17 times, every time one student as the presenter and others as participants to take part in the discussion experiments and therefore 4 second year's and one first year's master student gave presentation twice, and the other 7 graduate and undergraduate students gave presentation once, and all of their heart rate data during the presentation were measured, 247 Q&A segments were recorded.

3.2 Heart-Rate Data Analysis

We computed 18 HR and HRV features, e.g., mean, standard deviation (std.), and root mean square successive difference (RMSSD), from all Q&A segments as well as the separate question and answer periods. Which we gave more details in our previous work [2].

3.3 Human-Scoring Data Analysis

In this study, we asked all of the discussion participants to evaluate the answer-quality of Q&A segments. And there were 112 Q&A segments were evaluated as low-quality and 135 Q&A segments were evaluated as high-quality. We firstly tried to investigate that if the evaluation scores have difference between the question-askers and the other participants. Cohen's Kappa [3] was adopt to measure the interrater agreement between different items. We obtained a kappa value of 0.67, which means the question-askers had substantively the same evaluation opinions regarding the answer-quality with other participants. We then attempt to check if the difficulty level of questions would affect the quality of answers. However there were only 7 questions of Q&A segments were considered "Difficult" to answer. No questions were considered "Very Difficult" to answer. Therefore, we believe that among the data we collected, there is no need to consider the Q&A segments' question difficulty level at this stage

3.4 Machine-Learning Method for Evaluation

In our recognition task, we experimented with different types of classifiers including logistic regression, support vector machine, and random forest. We performed leave-one-student-out cross validation and reported mean AUC scores to evaluate our models overall recognition abilities. We also reported mean recall scores based on low-quality-answers to check our models' ability regarding recognition of low-quality answers as shown in Table 1.

Table 1: Mean AUC scores of models and mean recall scores based on low-quality answers

Model	Mean AUC	Mean recall scores based on low-qulaity answers
LR	0.76	0.69
SVM	0.77	0.65
RF	0.79	0.78

From our results reported in Table 1, for our answer-quality evaluation task, we got well mean AUC scores for all of the models, especially the RF classification model. This suggests that question-askers' heart rate data could be used to evaluate their answer-quality in lab discussion experiments. In addition, we also reported mean recall scores based on low-quality answers recognition task, for the same student Q&A data, RF could better recognize their low-quality answers than the other two models. Furthermore, LR model presents a stronger recognition ability than SVM model regarding the low-quality answers recognition problem even though with a slightly lower overall classification performance. In our leave-one-student-out cross-validation evaluation process, we also took insight into each students' answer-quality recognition results, for the master students who have a higher level experience of discussion presentation, there were more false negative recognition samples which means the answers that question-askers evaluated high-quality, but students presented a low-confidential mental states when giving answers and so that our classifiers recognized it as low-quality answers. We took an interview with these students and also investigated these answer statements and found that, the master students have more skills and ability to give answers seems like correct to question-askers even though they don't have confidence to answer them, we could also find some special words in these answers' statements such as "maybe", "perhaps" exhausted and this is also indicated a low-confidential mental states, so our heart rate models could recognized it as low-quality cases but not human evaluation models. We are thinking about these kinds of low-quality are easy to be ignored by human but using our HR-models we can recognize them precisely.

Conclusion

In our study, we analyzed 12 students, 17 times' discussion experimental data, including their heart rate and Q&A segments. Machine learning models were adopt to predict the answer-quality of Q&A segments, and the results confirmed our hypothesis that students' HR data could be used to evaluate their answer-quality.

Reference

[1] K. Nagao, K. Inoue, N. Morita and S. Matsubara. Automatic Extraction of Task Statements from Structured Meeting Content. Proc. of the 7th International Conference on Knowledge Discovery and Information Retrieval, pp.307-315, 2015.

[2] S. Peng and K, Nagao. Automatic Evaluation of Presenters' Discussion Performance Based on their Heart Rate. In Proceedings of the 10th International Conference on Computer Supported Education (CSEDU 2018), Vol. 2, pp. 27-34, 2018.

[3] J.R. Landis, G.G. Koch. The Measurement of Observer Agreement for Categorical Data. Biometrics. 33 (1): 159-174. doi:10.2307/2529310. JSTOR 2529310. PMID 843571.1997.