Effective Diagnostic Feedback for Online Multiple-Choice Questions

Ruisheng Guo, Dominic Palmer-Brown, Sin Wee Lee, and Fang Fang Cai

Faculty of Computing, Londonmet University, 166-220 Holloway Road, London N7 8DB
{r.guo1,ff.cai,d.palmer-brown}@londonmet.ac.uk,
s.w.lee@uel.ac.uk

Abstract. When students attempt MCQs (Multiple-Choice Questions) they generate invaluable information which can form the basis for understanding their learning behaviours. In this research, the information is collected and automatically analysed to provide customized, diagnostic feedback to support students' learning. This is achieved within a web-based system, incorporating the SDNN (Snap-drift neural network) based analysis of students' responses to MCQs. This paper presents the results of a large trial of the method and the system which demonstrates the effectiveness of the feedback in guiding students towards a better understanding of particular concepts.

Keywords: learning behavior, diagnostic feedback, neural networks, on-line multiple-choice questions.

1 Introduction

In recent years, e-learning has become commonplace in higher education. The involvement of intelligent e-learning systems has the potential to make higher education accessible with increasing convenience, efficiency and quality of study. According to the National Student Survey reports (2007- 2010) [1], in England, only about half of students believe that: 1, feedback on their work has been prompt; 2, feedback on their work has helped to clarify things they did not understand; 3, they have received detailed comments on their work. These reports reveal that the feedback and its related fields are one of the weakest areas in higher education in England. This research investigates the relative effectiveness of different types of feedback, and how to optimize feedback to facilitate deep learning. It compares and contrasts several methods in order to investigate the effectiveness of using intelligent feedback towards modeling the stages of students' knowledge. The investigation will lead to an understanding of the potential of the on-line diagnostic feedback across different subject areas.

The Virtual Learning Environment presented in this paper provides a generic method for intelligent analysis and grouping of student responses that applicable to any area of study. This tool offers important benefits: immediate feedback, significant time-saving evaluating assignments, and consistency in the learning process. The time taken to create the feedback is well spent not only because this feedback can be reused, but also it is made available through the system to large numbers of students.

L. Iliadis et al. (Eds.): AIAI 2012, IFIP AICT 381, pp. 316–326, 2012.

2 Background and Review of Previous Work

Rane and Sasikumar [2] pointed out that, intelligent tutoring systems attempt to simulate a teacher, who can guide the student's study based on the student's level of knowledge by giving intelligent instructional feedback. According to Blessing, Gilbert, Ourada and Ritter [3], the intense interaction and feedback achieved by intelligent tutoring systems can significantly improve student learning gains. To make the feedback effective and meaningful, a range of quality attributes need to be achieved. Hatziapostolou and Paraskakis [4] summarized the work by Race (2006), Irons (2008) and Juwah et al (2004) and suggested that in order to improve learning gains, formative feedback should address as many as possible of the following attributes, including constructive [5], motivational [4], personal [6], manageable [4], timely [7] and directly related to assessment criteria and learning outcomes [8].

Many researches investigating the effect of different types of feedback in web-based assessments showed the positive results of using MCQs in online test for formative assessment (e.g. [9] [10] [11]). Higgins and Tatham [9] studied the use of MCQs in formative assessment in a web-based environment using WebCT for a level 1 unit on undergraduate law degree. They summed that they could forecast all the possible errors for a question and write a general feedback for this question. However, using this type of feedback, it could be difficult to predict all the possible errors and produce the general feedback for a combination of questions, and it would be impossible for a large test banks (e.g. 3 questions with 5 answers would require 125 answer combinations; 5 questions with 5 answers require 3125 combinations, etc.). Payne et al [11] assessed the effectiveness of three different forms of feedback (corrective, corrective explanatory, and video feedback) used in e-learning to support students' learning. This type of feedback shows exactly which questions are answered correctly or not, with further corrective explanation and video feedback. Our approach to feedback is different from the above. The intelligent diagnostic feedback we present is concept-oriented instead of question-oriented. The learners are encouraged to review the concepts they misunderstood through the feedback in order to retake the test again and study further. It is important that each category of answers is associated with carefully designed feedback based on the level of understanding and prevalent misconceptions of that category-group of students so that every individual student can reflect on his or her learning level and certain mistakes using this diagnostic feedback. In addition, when students retake the test they receive new feedback according to his or her knowledge state, which in turn leads to more self-learning. Moreover, concept-based feedback can also prevent the student from guessing the right answers; if the students do not read the diagnostic feedback carefully, they may not even know which questions were answered incorrectly. According to our current research, there are no other reported studies on MCQs and formative web-based assessment which have used any similar form of using intelligent agent to analyse the students' response in order to provide diagnostic feedback.

3 Multiple-Choice Questions Online Feedback Systems (M-OFS)

To analyse the students' answers, and integrate over a number of questions to gain insights into the students' learning needs, a snap-drift neural network (SDNN) approach

is proposed. SDNN provides an efficient means of discovering a relatively small and therefore manageable number of groups of similar answers. In the following sections, an e-learning system based on SDNN is described.

3.1 Snap-Drift Neural Networks (SDNNs)

One of the strengths of the SDNN is the ability to adapt rapidly in a non-stationary environment where new patterns are introduced over time. The learning process utilises a novel algorithm that performs a combination of fast, convergent, minimalist learning (snap) and more cautious learning (drift) to capture both precise sub-features in the data and more general holistic features. Snap and drift learning phases are combined within a learning system that toggles its learning style between the two modes. On presentation of input data patterns at the input layer F1, the distributed SDNN (dSDNN) will learn to group them according to their features using snap-drift (Lee et al., [12]). The neurons whose weight prototypes result in them receiving the highest activations are adapted. Weights are normalised weights so that in effect only the angle of the weight vector is adapted, meaning that a recognised feature is based on a particular ratio of values, rather than absolute values. The output winning neurons from dSDNN act as input data to the selection SDNN (sSDNN) module for the purpose of feature grouping and this layer is also subject to snap-drift learning.

The learning process is unlike error minimisation and maximum likelihood methods in MLPs and other kinds of networks. These perform optimization for classification or equivalents by for example pushing features in the direction that minimizes error, without any requirement for the feature to be statistically significant within the input data. In contrast, SDNN toggles its learning mode to find a rich set of features in the data and uses them to group the data into categories. Each weight vector is bounded by snap and drift: snapping gives the angle of the minimum values (on all dimensions) and drifting gives the average angle of the patterns grouped under the neuron. Snapping essentially provides an anchor vector pointing at the 'bottom left hand corner' of the pattern group for which the neuron wins. This represents a feature common to all the patterns in the group and gives a high probability of rapid (in terms of epochs) convergence (both snap and drift are convergent, but snap is faster). Drifting, which uses Learning Vector Quantization, tilts the vector towards the centroid angle of the group and ensures that an average, generalised feature is included in the final vector. The angular range of the pattern-group membership depends on the proximity of neighbouring groups (natural competition), but can also be controlled by adjusting a threshold on the weighted sum of inputs to the neurons. The output winning neurons from dSDNN act as input data to sSDNN module for the purpose of feature grouping and this layer is also subject to snap-drift learning.

3.2 Training Neural Network

The E-Learning Snap-Drift Neural Network (ESDNN) is trained with the students' responses to questions on a particular topic in a course. The responses are obtained from the previous cohorts of students. Before training, each of the responses from the students is encoded into binary form in preparation for presentation as input patterns for

ESDNN. Table 1 shows examples of a possible format of questions for five possible answers and some encoded responses. This version of ESDNN is a simplified unsupervised version of the snap-drift algorithm (Lee et al., [12]) as shown in Fig. 1.

Table 1. Example of input patterns for ESDNN

Codification	A:00001	B:00010	C:00100	D:01000	E:10000	N/A:00000
Response	Recorded Response					
[C,D,B,A]	[0,0,1,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,1]					
[E,B]	[1,0,0,0,0,0,0,0,1,0]					

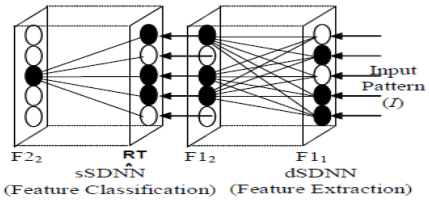

Fig. 1. E-learning SDNN architecture

During training, on presentation of an input pattern at the input layer, the dSDNN will learn to group the input patterns according to their general features. In this case, 5 F12 nodes, whose weight prototypes best match the current input pattern, with the highest net input are used as the input data to the sSDNN module for feature classification. In the sSDNN module, a quality assurance threshold is introduced. If the net input of an sSDNN node is above the threshold, the output node is accepted as the winner; otherwise a new uncommitted output node will be selected as the new winner and initialised with the current input pattern. For example, for one group, every response might have in common the answer C to question 2, the answer D to question 3, the answer A to question 5, the answer A to question 6, the answer B to question 8, and the answer A to question 10. The other answers to the other questions will vary within the group, but the group is formed by the neural network based on the commonality between the answers to some of the questions (four of them in that case). From one group to another, the precise number of common responses varies in theory between 1 and X, where X is the number of questions. In this experiment, where there are 10 questions in 1st English trial (Section 5), the groups had between 5 and 8 (Trial 1) common answers. More details of the steps that occur in ESDNN and the ESDNN learning algorithm are given in (Lee et al., [12]). The training relies upon having representative training data. The number of responses required to train the system so that it can generate the states of knowledge varies from one domain to another. When new responses create new groups, more training data is required. Once new responses stop creating new groups, it is because those new responses are similar to previous responses, and sufficient responses to train the system reliably are already available. The number of groups formed depends on the variation in student responses.

3.3 How the System Guides Learning

The feedback is designed by academics so that it does not identify which questions were incorrectly answered. The academics are presented with the groups in the form of tem-plates of student responses. For example, "A/D B mix" represents a group characterized by all the students answering A or D to question 1, B to question 2, and mixed answers to question 3. Hence, the educator can easily see the common mistakes in the groups of the student answers highlighted by the tool. The feedback texts are associated with each of the pattern groupings and are composed to address misconceptions that may have caused the incorrect answers common to that pattern group. The student responses, recorded in the database, can be used for monitoring the progress of the students and for identifying misunderstood concepts that can be addressed in subsequent face-to-face sessions. The collected data can be also used to analyze how the feedback influences the learning of individual students by following a particular student's progress over time and observing how that student's answers change after reading the feedback. Student responses can also be used to retrain the neural network and see whether refined groupings are created, which can be used by the educator to improve the feedback. Once designed, MCQs and feedbacks can be reused for subsequent cohorts of students.

4 Approach

In order to evaluate the performance and effectiveness of this novel e-learning system, target-oriented testing of the system needs to be carried out in different fields. Furthermore, we also aim to enhance this system to overcome its deficiencies during practical applications. Thus, this study is composed of three main parts. Firstly, we evaluated the M-OFS system by collecting and analysing a large number of testing data reflecting the students' learning gains by using this system as well as the survey and interview data reflecting the students' satisfaction and attitudes towards this system. Secondly, the investigation leads to an understanding of the potential of the on-line diagnostic feedback approach across different subject areas. Thirdly, this research should also produce guidelines for the design principles of on-line MCQs in the context of diagnostic feedback learning environments. The details of this experiment which are conducted to assess the use of M-OFS during academic year 2010-2011 are reported below. Four hypotheses are formulated: (H1) students are satisfied with using M-OFS; (H2) students improved their understanding by reading given feedback; (H3) in a separate MCQs paper test, students get higher mark in the first test than the second trial by learning from the M-OFS; (H4) in the final examination, the average score of the experimental group is higher than the average score of the control group.This research used 5 instruments: 1, four previous MCQ test (1 conducted in 2008 and 3 in 2009) results are collected to train the ESDNN; 2, two separate MCQs paper tests are applied before and after the system trial; 3, compare and contrast the scores between first and last attempt during the system trial; 4, compare and contrast the final examination grades between experimental group and control group; 5, survey and interview to assess learner's satisfaction and motivation.

4.1 Data Collection

Data of Six trials were collected in total. It includes three English trials, two Math trials and one Plagiarism trial. In this paper, it will present the details of data collection of 1st English trial as below:

The data for training is collected from three previous year's MCQs tests (2008-2010). For these three tests, 94 students' answers were used to training. The trials data were collected during academic year 2010-2011. The data of two separate MCQ paper tests and final examination results were gathered. 83 students entered the survey and 16 students were randomly selected for interview. The states of knowledge of students were achieved by using ESDNN.

4.2 English Experiments

To investigate and evaluate how the M-OFS guide and support students to learn, three English experiments were under taken by level 2 and level 3 students at JinQiao University and Kunming Technology University in China during the academic year 2010-2011. The 1st experiment is introduced below.

In the first experiment, data was collected from 148 students taking English language courses whom were randomly separated into two groups. The experimental group of 83 students used M-OFS, and the control group of 65 students received the same training but without using M-OFS. The system trial includes 10 MCQs with 4 potential answers, related to English grammar. The duration of this trial is flexible. When students were using M-OFS, they were encouraged to answer the MCQs (submit their answers) as many times as they wish until they got all the correct answers or gave up (students were not given answers or how many answers were correct in their feedback, except that they answered all correct answers). Two MCQ paper tests with different questions from system trials were applied to 116 students, and 83 students participated in both paper test and system trial. 83 students completed survey after second paper test. System trial, paper test and survey were completed in practice lessons in computer room at JinQiao University.

5 Empirical Study

This section discusses the results from the first experiment in order to evaluate the effectiveness of using M-OFS to support students' deep learning.

The survey and interview were conducted after the system trial. 83 (100%) students conducted the survey. 16 (19%) students were randomly chosen for interview. For the survey, 71.1% students are satisfied with using system. 84.4% students think the feedback is what they need. Using M-OFS to learn were positively evaluated by students, illustrate that the hypotheses H1 is supported. 90.4% students would like to use the system again. 92.8% students would like to recommend the system to a friend or classmate in the future. 81.9% students have never used similar system before. For interviews, most students (94%) feel this system is useful and helps them to improve their knowledge, it indicates the hypotheses H1 is supported as well; moreover, 69% students want the exact answers in the feedback in the end. Students also want a picture

of their learning process which can point out their weakness and a suggestion of how to improve their English. Some students feel that if they tried many time but cannot find the correct answer, they will lose patience in the end.

5.1 Experiment and Result

148 students are involved in the first experiment. 116 students completed the separate MCQs paper test before and after using the system. 83 students participated in system trial, and separate MCQs paper tests.

For system trials, a total of 1118 answers/attempts were submitted and a total 2143 minutes were spent by 83 participants. All of the students submitted their answers at least once. The maximum number of attempts was 106 times and the minimum was 1. The average attempts for each student is 13.5 times. The average time spent by each student is 25.8 minutes and the average time of each attempt is 1.92 minutes. 2 students (2.4%) spent more than 60 minutes. 35 (42.2%) students spent more than the average time. No students achieved the all correct answers at the beginning. 55 (66.3%) students increased their scores by an average of 12.77%, whilst 1 student increased his score by 70%. In this trial, with 10 questions and 4 possible answers, there are more than 1 million possible combinations of answers, thus the students are unlikely to make improvement by guessing answers; hence, the results show the feedback had a positive impact which partially supports hypotheses H2.

For separate MCQs paper tests, the average score before system trial is 51.6%, and the average score after system trial is 59.15%. One student (Student no. 200916031222) increased his score by 40%. 74% students increased their scores. In this test, the students were not given any answers or feedback between first (before system trial) and second (after system trial) test; furthermore, the first trial were applied 3 hours before the system trial and the second test were conducted 30 minutes after system trial; hence, the students are only learnt by using M-OFS but not any other ways; thus the results above are confident, therefore partially supporting hypotheses H3. In addition, this result also can partially support hypotheses H2.

For final examination, both the experimental group and the control group enter the same 4 days final examination. The experimental group got 79.52% and control group got 71.28% in English grammar module. This result confirms the hypotheses H4; furthermore, it also supports hypotheses H2.

5.2 Some Group Behavioural Characteristics

Previous work has made an initial investigation of the behavioural characteristics of students during their learning interaction with a diagnostic feedback system [13]. In order to explore the characteristics of students, and relate these to student responses and performance in the tests, five behavioural variables were analysed: the number of attempts (submissions), the average score changed between attempts, the average score at the end of trial, the amount of time spent to make each attempt, and the learning duration. Fig. 2 illustrates a learning behaviour of this group of students by analysing the relationship between average scores increased and learning duration. Each blue point represents average scores increased of all students used the same learning time,

and its coordinate of x-axis represents student's learning duration, and its coordinate of y-axis represents average scores increased. It can be achieved from this figure that average scores increased when students spent more time on studying from the system.

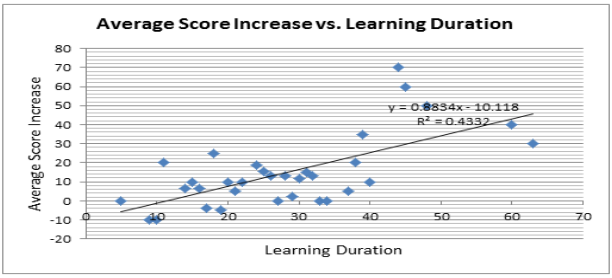

Fig. 2. Average Score Increased vs. Learning Duration

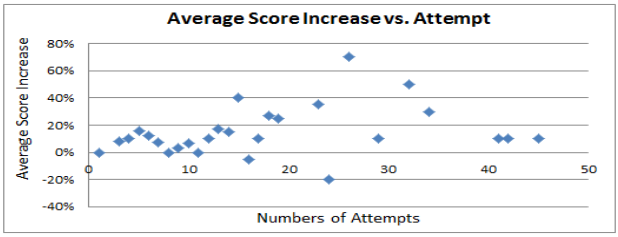

Fig. 3. Average Score Increase vs. Attempt

Fig. 3 illustrates a learning behaviour of this group of students by analysing the relationship between average scores increased and number of attempts. Each blue point represents average scores increased of all students did the same number of attempts, and its coordinate of x-axis represents number of students' attempts, and its coordinate of y-axis represents average scores increased. It can be achieved from this figure that average scores increased when students did more attempts before peak, and average scores no longer increased when students did 26 attempts, and average scores decreased by doing more attempts after peak.

Table 2. Short ***learning duration***: time spent on learning < 25.6 minutes; ***Long learning duration***: time spent on learning > 25.6 minutes

Behavioral Group	Average score changed	Average score at the end	Number of students
Short learning duration	9.17%	48.33%	48
Long learning duration	17.14%	61.43%	35

Table 3. Many ***attempts***: number of attempt >13.4, ***Few attempts***: number of attempt <13.4

Behavioral Group	Average score changed	Average score at the end	Number of students
Many attempts	20.77%	63.07%	26
Few attempts	8.77%	49.65%	57

Table 4. Slow ***attempt***: average time spent on each attempt >1.92 minutes; ***Rapid attempt***: average time spent on each attempt <1.92 minutes

Behavioral Group	Average score changed	Average score at the end	Number of students
Slow attempt	12.96%	53.52%	54
Rapid attempt	11.72%	54.48%	29

Table 5. Rapid ***few attempt***: average time spent on each attempt <1.92 minutes and number of attempt <13.4; ***Few slow attempts***: number of attempt <13.4 and average time spent on each attempt >1.92 minutes; ***Many rapid attempts***: number of attempt >13.4 and average time spent on each attempt <1.92 minutes; ***Slow many attempts***: average time spent on each attempt >1.92 and number of attempt >13.4

Behavioral Group	Average score changed	Average score at the end	Number of students
Rapid few attempts	2.22%	48.89%	9
Few slow attempts	10%	50%	48
Many rapid attempts	16%	57%	20
Slow many attempts	36.67%	81.67%	6

Table 6. Slow, ***few attempt and short learning duration***: average time spent on each attempt >1.92 minutes, number of attempt <13.4, and time spent on learning <25.6 minutes; ***Rapid, many attempts and short learning duration***: average time spent on each attempt <1.92 minutes, number of attempt >13.4, and time spent on learning <25.6 minutes; ***Rapid, many attempt and long learning duration***: average time spent on each attempt <1.92 minutes, number of attempt >13.4, and time spent on learning >25.6 minutes; ***Slow, few attempt and long learning duration:*** average time spent on each attempt >1.92 minutes, number of attempt <13.4, and time spent on learning >25.6 minutes

Behavioral Group	Average score changed	Average score at the end	Number of students
Slow, few attempt and short learning duration	9.35%	48.71%	31
Rapid, many attempts and short learning duration	15.7%	54.29%	7
Rapid, many attempt and long learning duration	16.15%	58.46%	13
Slow, few attempt and long learning duration	11.18%	52.35%	17

Many attempts and long learning time are consistently associated with good score increases, and hence represent successful learning strategies amongst the students.

6 Summary

Six trials in three totally different subject areas have been carried out in three universities in two countries which are the UK and Chinese: 3 English trials, 2 Mathematics trials and 1 Java Programming trial. The English trials are very successful

with 500 students participated. A large volume of data has been captured during the trials. The results of the first English system trials show that the average score is increased by 12.8% at the end. The results of the separate MCQ paper test present the average mark of the group is increased by 7.6%. In addition, the final examination, the average mark of experimental group is 8.2% higher than the control group. Furthermore, student surveys show that 71.1% students are satisfied with our e-learning system and 84.4% students feel that the intelligent diagnostic feedback is what they need.

7 Conclusion and Future Work

In this paper, a novel method for using snap-drift in a diagnostic tool to provide intelligent diagnostic feedback is presented. There are several innovative features of the work: this is the first time that the neural network diagnostic feedback approach in MCQ has been systemically applied to large cohorts of students and evaluated across a range of different subject areas; and the use of a neural network to discover groups of similar answers that represent different knowledge states of the students. The feedback targets the level of knowledge of individuals, and their misconceptions, guiding them toward a greater understanding of particular concepts. The results of the experiment demonstrate that an improvement in the learning process can be achieved.

In future work, it is intended to compare the effects of the feedback to the effects of other types of feedback already studied in the literature. Another promising avenue for further investigation is the extension of the tool to support knowledge state transition diagram construction and statistical data collection, which could help instructors to analyze the difficulty of the MCQs and to track students though the developmental stages of their learning.

References

1. Hefce national student survey, HEFCE, London, U.K. (2007-2010), http://www.hefce.ac.uk/learning/nss/
2. Rane, A., Sasikumar, M.: A constructive learning framework for language tutoring. In: Iskander, M. (ed.) Innovations in e-Learning, Instruction Technology, Assessment, and Engineering Education. Springer, The Netherlands (2007)
3. Blessing, S., Gilbert, S., Ourada, S., Ritter, S.: Lowering the bar for creating model-tracing intelligent tutoring systems. In: Luckin, R., et al. (eds.) Artificial Intelligence in Education. IOS Press (2007)
4. Hatziapostolou, T., Paraskakis, I.: Enhancing the Impact of Formative Feedback on Student Learning through an Online Feedback System. Electronic Journal of e-Learning 8(2), 111–122 (2010), http://www.ejel.org
5. Nelson, M.M., Schunn, C.D.: The Nature of Feedback: How Different Types of Peer Feedback Affect Writing Performance. Instructional Science 37(4), 375–401 (2009)
6. Garber, P.R.: Giving and Receiving Performance Feedback. HRD Press, Canada (2004)
7. Race, P.: The Lecturer's Toolkit – A Practical Guide to Assessment, Learning and Teaching, 3rd edn. Routledge, London (2006)

8. Springgay, S., Clarke, A.: Mid-Course Feedback on Faculty Teaching: A Pilot Project. In: Darling, L.F., Erickson, G.L., et al. (eds.) Collective Improvisation in a Teacher Education Community, ch. 13, pp. 171–185. Springer, The Netherlands (2007)
9. Higgins, E., Tatham, L.: Exploring the potential of multiple choice questions in assessment. Learn. Teach. Action 2(1) (2003)
10. Kuechler, W.L., Simkin, M.G.: How well do multiple choice tests evaluate student understanding in computer programming classes? J. Inf. Syst. Educ. 14(4), 389–399 (2003)
11. Payne, A., Brinkman, Wilson, F.: Towards effective feedback in e-learning packages: The design of a package to support literature searching, referencing and avoiding plagiarism. In: Proceedings of HCI 2007 Workshop: Design and Use and Experience of e-Learning Systems, pp. 71–75 (2007)
12. Lee, S.W., Palmer-Brown, D., Draganova, C.: Diagnostic Feedback by Snap-drift Question Response Grouping. In: Proceedings of the 9th WSEAS International Conference on Neural Networks (NN 2008), pp. 208–214 (2008)
13. Alemán, J.L.F., Palmer-Brown, D., Jayne, C.: Effects of Response-Driven Feedback in Computer Science Learning. IEEE Trans. Education 54, 501–508 (2011)