Peer review to improve artificial intelligence teaching

Raquel M. Crespo García, Julio Villena Román, and Abelardo Pardo
Universidad Carlos III de Madrid, Spain
{rcrespo, jvillena, abel}@it.uc3m.es

Abstract – Using a team-work, project-based methodology is a common approach when teaching Artificial Intelligence. However, a major drawback of such approach is that AI courses comprise a wide syllabus composed of quite independent topics. In consequence, students focus on one single topic from the entire course contents: although deep learning of such topic is probably ensured, learning of the rest of the contents is also probably much more superficial. In this paper, peer review is proposed as a complement to project-based learning. Students work not only on their project about a chosen topic, but also review peers’ projects on distinct topics, providing them with a wider comprehension of the global syllabus of the course. Empirical results of the application of this approach to an actual course in Artificial Intelligence for senior students in Telecommunication Engineering are presented too. Analysis focuses on the effects of the reviewing task, as it is the one which broadens students learning. Positive results have been achieved, thus validating the interest of peer review as a useful instrument for learning improvement.

Index Terms – Peer review, Artificial Intelligence.

INTRODUCTION

Teaching an introductory course on artificial intelligence for undergraduate students brings up interesting challenges, in which the high motivation of students plays a main role. However, the usual drawback in these courses is the diversity and complexity of the different topics included in the course, usually including knowledge representation, problem solving, learning algorithms, fuzzy logic, inferential systems, agents, etc.

Working on a course project is a common approach for subject learning and evaluation purposes. However, the intrinsic features of artificial intelligence courses raise specific problems to this approach. Students focus on the project topic, exploring it in depth, but only have a superficial look at the other topics of the course. Proposing specific projects for each topic is not a realistic option, as time resources are very limited for both students and teaching staff.

In the educational context, peer assessment is defined as an arrangement in which individuals consider the amount, level, value, worth, quality, or success of the products or outcomes of learning of peers of similar status [1]. Based on social constructivism, the benefits as well as a detailed topology of peer review in education have been carefully studied in several publications (see [1] for an excellent survey). A number of supporting programs for peer review in educational environments are reported since 1995; more recent tools such as CPR [2], PG [3], NetPeas [4], OPAS [5] or OASIS [6] use the web to manage peer interaction. And applications of peer review in education, as well as analysis of its reliability and validity, are reported in a number of technical and non-technical educational contexts, including computer science [6], operating systems [7]-[10] and computer architecture [11]. Here, Artificial Intelligence reveals itself as an ideal scenario for this methodology.

In this paper, peer review is proposed as a complement to the project-based learning approach, to overcome its practical limitations and drawbacks in the context of artificial intelligence courses. Peer review is suggested in order to allow students to explore in depth other topics besides the one developed in their own projects. Critical analysis of how the methodology affects the learning experience is done, considering both subjective and objective results. While emphasis is traditionally put on peer feedback benefits and reliability of students as evaluators, this paper focuses on the effects of the evaluation task itself (as it is the task which broadens students’ learning to other topics beyond their own one).

CONTEXT DESCRIPTION

A typical undergraduate course on artificial intelligence (AI) consists on an introductory course covering a wide diversity of topics. Communications Networks Intelligence (CNI) is one of such courses: an elective course for senior Telecommunication Engineering students, in which the objective is to analyze the concept of intelligence in computers and communications systems and to study the available techniques for integrating these functionalities into existing networks and services. At the end of the course, students are expected to know the basics of AI and some applications and fields where these technologies can achieve most significant progress. Table I summarizes the contents of the course [12].

Theoretical lectures are combined with practical sessions at laboratory. At the end of the course, students must develop a compulsory project to show that they are able to design and implement an intelligent system which solves concrete aspects of engineering. This project is chosen among a list of proposals and done in groups.
TABLE I
SYLLABUS OF THE COURSE

- Artificial Intelligence concepts
  - Knowledge representation
  - Formal logic basics and inferential systems
  - Management of uncertainty
  - Problem solving and searching strategies
  - Learning and knowledge acquisition

- Artificial Intelligence applications
  - Knowledge-based systems
  - Agents
  - Linguistic engineering and information retrieval
  - Data mining
  - Intelligent systems and communication systems

The number of students enrolled in the course has been continuously increasing during the last years, as shown in Figure 1. Although this growth is a handicap for personalized teaching, it brings up interesting possibilities for introducing collaborative learning in the course.

![Graph showing the number of enrolled students from 2002/03 to 2004/05](image)

FIGURE 1
NUMBER OF STUDENTS ENROLLED IN CNI.

Annual surveys for teacher evaluation [13] show that the main reasons for choosing the course are the appealing syllabus and express recommendation from students of previous years. Although the high passing rate could also influence the choice, we think that it doesn’t constitute a key factor as this rate is similar to other elective courses. Surveys also report a very positive evaluation of the course.

MOTIVATION

Despite the highly satisfactory results, some matters of concern recurrently arise every year, including the following aspects:

- How to avoid a biased viewpoint of the subject? Interest of the topics is not discussed, as students themselves think. However, they are quite different and independent. As shown in Table I, the syllabus includes many techniques and applications in different areas. Students sometimes find it very difficult to reinforce those basic concepts or even to learn them. Although the schedule has been rearranged to solve some of these problems (mainly eliminating those secondary topics which are less interesting for engineering students), there is not much room for improvement. To make things worse, the final project must necessarily focus only on one or two aspects, due to time and effort constraints.

- How to increase students’ motivation? Despite its great attraction, this course is elective and competes against other courses (compulsory, more important, failed from previous years…) to catch the student attention and effort.

- How to promote the development of critical non-technical abilities? Each year, lacks are detected in some basic competences, such as communication skills (oral communication, writing skills), critical analysis, self-confidence, etc. In short, abilities which are complementary to the technical aspects but have at least the same importance in the training of future professionals.

Peer review methodology seems to suit well in this context. The main idea is to use the active and proactive participation of the student as a booster for learning, both for technical aspects and also for communication skills. At the same time, active and proactive participation relies on the student motivation, as no motivation means no participation. Formally, the main objectives for improvement are:

- Quality learning: reinforce the student knowledge and study course topics in depth, promoting continuous work.

- Complete learning: broaden the student focus, complementing his/her training and to go into other topics.

The first point, learning in depth, is mainly addressed by the development of the final project. The effectiveness of this methodology is unquestionable and it has achieved really positive results in this particular course. The second point, broaden the students’ learning beyond the topic of their own project, is the main objective of the application of the peer review methodology in this context.

Additionally, issues for potential improvement include:

- Integral education: promote the development of essential non-technical abilities such as team work, communication skills and critical analysis.

- The student in the center of the educational process: promote a role in which students are more active, proactive, autonomous and responsible in the whole process.

- Improve the evaluation process: help students whose measured results are worse than their actual abilities due to external factors such as time pressure, insecurity, etc., allowing them to practice in a friendlier environment, thus increasing their self confidence before the actual exam.

COURSE DEVELOPMENT

During the 2004-05 course, student participation was included as a core concept during the classes: for each class, one or two students could voluntarily participate with a brief presentation about aspects of the topic for the day. Some objectives were covered: include some complementary aspects which are usually left apart, develop communication skills (slides preparation and speech), and make classes more dynamic. The
initial idea was to make one presentation in each class, but the high participation forced to change the schedule to include two monographic round tables to give the opportunity to participate to more students. All students evaluated this initiative as very positive.

With respect to the final project, topics which were finally chosen by students are shown in Table II. Each group had to submit a written report about their work, describing their system, solution and techniques (and source code, if any) and do a brief oral presentation in the classroom. Both the presentation and the written report were evaluated using the peer review methodology.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Project</th>
<th>#Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation</td>
<td>Simulator: Robocode</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Simulator: CodeRuler</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Simulator: RARS</td>
<td>6</td>
</tr>
<tr>
<td>Data Mining</td>
<td>DM: Meteorological prediction</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>DM: Diagnosis in cardiology</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Face recognition with NN,NN.</td>
<td>2</td>
</tr>
<tr>
<td>Expert Systems</td>
<td>Expert system for tourism</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Expert system &quot;Car dealer&quot;</td>
<td>2</td>
</tr>
<tr>
<td>Other</td>
<td>AI and sustainable development</td>
<td>2</td>
</tr>
</tbody>
</table>

On one hand, evaluation of oral presentations was done by teams, using a questionnaire focused on both technical and communication aspects. Each group had to evaluate all other groups, so they had to attend every presentation session. Rankings were requested, besides absolute scores, in order to force students to take decisions.

On the other hand, reports evaluation was done individually. Each student was given four projects to review: one about the same topic and application as his/her project, another in the same topic (same technique but different application), another one in a completely different topic, and finally his/her own project (self-evaluation). The revision questionnaire is shown in Table III. The project-reviewer allocation was done applying a pseudo-random strategy to ensure that each student would not evaluate the same projects as his/her team peers.

<table>
<thead>
<tr>
<th>Score from 1 (worst) to 5 (best):</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The work is dealt with adequate depth. Detail aspects too much emphasized and aspects that would need more effort</td>
</tr>
<tr>
<td>2. Global evaluation of work</td>
</tr>
<tr>
<td>3. Grammar/spelling mistakes (5=none to 1=more than 10)</td>
</tr>
<tr>
<td>4. Clarity: the writing is correct (clear, comprehensible...)</td>
</tr>
<tr>
<td>5. Style: the language is appropriate (scientific language, formal, avoids colloquial expressions...)</td>
</tr>
<tr>
<td>6. Semantics: the work done is correctly described</td>
</tr>
<tr>
<td>7. Global evaluation of writing (1=bad to 5=excellent)</td>
</tr>
<tr>
<td>8. The described hypotheses, process and results are correct</td>
</tr>
<tr>
<td>9. Evaluate the oral presentation</td>
</tr>
<tr>
<td>10. Global evaluation of report</td>
</tr>
<tr>
<td>11. Have you learned anything by reviewing this report? (from 1=nothing at all to 5=lot)</td>
</tr>
<tr>
<td>12. My knowledge of the course (1=null to 5=very high)</td>
</tr>
<tr>
<td>13. My interest in the course (from 1=null to 5=very high)</td>
</tr>
</tbody>
</table>

RESULTS

Figure 2 shows the estimated time spent on reviewing each project. The graphic shows average values, but the standard deviation is rather low and similar for all cases. Thus, it can be assumed that answers are rather accurate. According to the data, the revision task demands a significant amount of time, considering that each student reviewed three projects plus his/her own. This is a fact to be considered when planning these activities.

Figure 2 also shows that difficult projects, according to students’ opinion in the evaluation questionnaire, demand more time for review, which seems sound, considering the greater length of the reports. The case of the project about “face recognition using neural networks” (FR-NN in the figure) is significant because reviewers spent almost twice as much time as the rest. This is due to the great interest in the project declared by the reviewers in their comments in the evaluation report.

For evaluation purposes, questions about each frequently selected project (Robocode, CodeRuler and RARS – simulation-, and meteorological prediction and diagnosis in cardiology –data mining–) were introduced in the written exam. In order to improve the test reliability, questions were inserted among the other exam questions (a total of 50 true/false questions), so that students thought that all of them counted equally for the final grade (though, in fact, they were not considered for the final grade). It is important to emphasize the relative empirical rigor of this kind of social experiments, with such small populations, whose behavior may be biased, and limited set of control questions. As a consequence, conclusions are more qualitative than quantitative, and should be interpreted in a general way.

In the following figures, “authors”/“non authors” are the students who did/did not develop the project, “reviewers” are the students who did not develop the project but evaluated it,
and “rest” stands for the students who neither developed the project nor evaluated it.

Figure 3 presents the average number of correct answers for each subset of questions, differentiating the results for authors, non-authors and both. For all topics, the average number of correct answers is higher for the authors than for the rest of the class, which is itself higher than the average for the students who had not chosen the project about that topic. These results confirm that working on a project in fact has its impact on learning.

![Graph showing average number of correct answers per topic]

Results also show surprisingly low results for questions about diagnosis in cardiology (data mining), even for the authors of projects on that topic. This may be explained because one of the questions turned out to be ambiguous and confusing for the students (this was discovered when evaluating the results).

Figure 4 shows a comparison of the average number of correct answers for authors, reviewers and the rest of students (neither authors nor reviewers). In all topics, the highest value corresponds to the authors. But reviewers’ results are themselves better than the results of the rest of students, except for the RARS project. This contradictory data can be explained by the low number of students assigned as reviewers for this project (this project was chosen only by two groups and so the number of reviewers was lower than for other topics), which restricts its statistical significance. These valuable results (again, considering the relative statistical relevance which is intrinsic to this kind of experiments with small populations) allow us to infer the validity of the peer review methodology as a learning instrument, i.e., students learn by reviewing other students’ projects. The difference between authors and reviewers is expected, considering the time spent reviewing compared to developing the project.

Figure 5 and Figure 6 show an interesting result relating to the confidence degree of students about their answers. First figure shows the number of correct/incorrect/blank answers for authors (that is, students asked about their own project). The second one shows the same data but for reviewers. Considering that students thought that incorrect answers had a penalty in the exam grade, we can assume that non-blank answers correspond to questions which the students were convinced to know.

![Graph showing percentage of correct, incorrect and blank answers (authors)]

![Graph showing percentage of correct, incorrect and blank answers (reviewers)]

We observe that authors had a really low number of blank answers (even none in some case -RARS-). Authors answered most of the questions about their own project, which is
indicative that they were confident about their topic (independently that they were indeed right or failed). However, reviewers were not so confident in their answers (the number of blank answers increases notably), especially for projects considered as most difficult (about data mining).

**STUDENTS FEEDBACK**

Overall, students have reported a positive evaluation of the experience. An open question was posed to the students in order to get their opinions about the peer review process, as well as positive and negative aspects they had found. In this section, a summary of the received concerns and conclusions is provided.

A significant part of the students did not understand the importance of a critical analysis of their own work:

— “Self-evaluations are completely useless”.

Students were reticent to evaluate their peers. Evaluation was often seen as a complex and demanding task and students did not feel themselves qualified and objective enough:

— “I can evaluate a work similar to mine, but I don’t know how to evaluate one about another topic”.

Thus, they would have preferred to delegate the responsibility of evaluation to the teaching staff (seen as “experts” on the subject):

— “The teacher is who can judge best. It is unfair that students evaluate, because they don’t know everything”.

But most students evaluated positively the experience, even considering it a demanding task, and reported learning improvements from the process:

— “I think that too much time is required for reading and commenting all the projects, but even so I think it is a good idea, because it allows us to realize which errors we usually make”.

Finally, some opinions reflected a mature analysis of the pros and cons of the process found:

— “The process of evaluation when the topic is unknown and, moreover, the authors are your own colleagues usually is quite contradictory morally. It is difficult to evaluate objectively your own peers (because you know them, they can be your friends, etc.). Self and peer evaluation of projects about the same topic as yours, allows you to realize about aspects which you had discarded from the beginning”.

**CONCLUSIONS**

Teaching Artificial Intelligence using a team-work, project-based methodology is a common approach. However, artificial intelligence courses usually comprise a wide syllabus, composed of quite different and independent topics. As a consequence, a major drawback of such didactic approach is that students focus on one single topic from the entire course contents. Although deep learning of such topic is probably ensured, learning of the rest of the contents is also probably much more superficial.

The application of peer review methodology has proved to be useful to overcome this limitation: results of an objective empirical student evaluation show that students effectively learn by reviewing other students’ projects. Regardless the limited statistical relevance of the experiment, we consider that these qualitative conclusions discover new teaching strategies to reach our initial objectives for improving the learning process.

Of course, there have been problems and difficulties related to the practical application of peer evaluation in the real classroom:

- One of the most frequently reported problems is the drop rate: some students do not submit their assigned reviews in time or they do not submit them at all.
- Anonymity is difficult to guarantee in small groups. Students often recognize their peers’ work, so the review is not anonymous and can inhibit objectivity.
- Another difficulty derives from the schedule rearrangements to achieve a better integration in the course plan and the student agenda. The sequential methodology of peer review process in which each phase depends on the result of the previous one (development → review → evaluation of reviews) is strongly sensitive to delays, which accumulate progressively. In consequence, reviews were done near the end of the course, when students have less time due to the proximity of the examinations, which affects to quality of some works.

Regarding the acquisition of non-technical abilities, it is very difficult to measure the impact of this experience, as it is more a continuous holistic process. Anyway, the exhaustive evaluation of both the presentation and the written report and consciousness of being evaluated by their friends pushed students to put a special attention to their work and, in general, it was better than in other occasions, according to the teaching staff opinion.

However although the effectiveness of the methodology as a learning instrument has been endorsed, its application as a grading instrument remains questionable. In this context, students are reticent to grade their peers, due to factors beyond the evaluation process itself. Thus, grade homogeneity and over marking are two major drawbacks of peer evaluation, as shown in other peer review case studies. As a conclusion, from the resulting data, peer review must be complemented with strict quality control mechanisms in order to guarantee its reliability as a grading instrument.

As a result from this experience, peer review will be integrated in the course in future editions and will also be expanded to other courses with similar concerns.

**ACKNOWLEDGMENT**

This work was partially funded by Programa MEC-CICYT project MOSAIC-LEARNING TSI2005-08225-C07-01/02 and Programa de Apoyo a Experiencias de Innovación Docente (Support Program for Experiences on Teaching Innovation), Universidad Carlos III de Madrid. The authors would also like to thank the students for their participation and involvement.

Session T1A

October 28 – 31, 2006, San Diego, CA
REFERENCES


