Evaluation of a Graduate Level Data Mining Course with Industry Participants

Peter Christen

Department of Computer Science, The Australian National University
Canberra ACT 0200, Australia
Email: peter.christen@anu.edu.au

Abstract

Data mining courses are increasingly being taught at many universities at both undergraduate and graduate levels. This paper reports on a new graduate level data mining course run for the first time in 2007 at a major Australian university. The course had almost 20% enrolments of industry based participants from both private and public sector organisations. This paper discusses the student population and presents the course structure and assessment. An empirical evaluation of student responses, conducted at the end of the course, is then provided, with an emphasis on differences in responses from graduate students and external participants. To the best of the author's knowledge, this is the first such detailed empirical evaluation of a data mining course.

Keywords: Data mining education, postgraduate studies, course evaluation, industry.

1 Introduction

With many businesses, government organisations and research projects collecting massive amounts of data, the techniques collectively known as *data mining* have in recent years attracted interest from both industry and academia. As a result, there is now an increasing demand from both industry and government agencies for graduates with data mining skills, and data mining courses are now being taught at many universities throughout the world.

Data mining is multi-disciplinary and draws from fields such as statistics, machine learning and AI, database technology, algorithms and data structures, high-performance and parallel computing, visualisation, and privacy and security technologies (Han and Kamber 2006). This wide spectrum challenges the teaching of data mining, as necessarily some prior knowledge in some of the above mentioned disciplines is required. As reported elsewhere (Saquer 2007), one major challenge when teaching data mining is that students will have different backgrounds, and likely have knowledge in some of the core disciplines of data mining. The wide range of concepts and techniques related to, and required by, data mining also limits either the depth or breath of how much can be covered in a normal one-semester university course.

The aim of the new graduate level course discussed in this paper was to cover more than just the core data mining techniques and algorithms (like classification, prediction, clustering and association rule mining), but to also expose students to other important issues relevant to the knowledge discovery in databases (KDD) process, ranging from data quality, pre-processing and integration to privacy and social impacts of data mining. Additionally, a second focus of the course was to give students insight into current data mining research through reading papers and an oral presentation of a selected research paper.

The course was not only advertised as a graduate level course on relevant university Web sites and student handbooks, but also announced to several local e-mail lists containing contacts from government agencies and private sector organisations known to have interests related to data mining. The course syllabus as available in the student handbook was:

> Large amounts of data are increasingly being collected by public and private organisations, and research projects. Additionally, the Internet provides a very large source of information about almost every aspect of human life and society.
>
> This course provides a practical focus on the technology and research in the area. It focuses on the algorithms and techniques and less on the mathematical and statistical foundations.

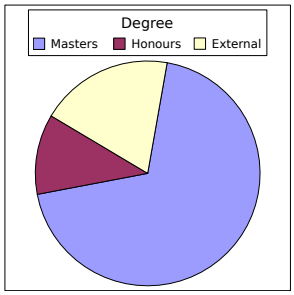
In the following section a short overview of related work is provided. In Section 3 some details about the student population is discussed, and in Sections 4 and 5 the course structure and assessment, respectively, are presented. Section 6 then contains a detailed discussion of the course evaluation based on an end-of-semester questionnaire and observations by the author. Finally, the paper is concluded in Section 7 with a discussion of potential changes and improvements for future offerings of this course.

2 Related Work

Only a very small number of reports that describe and evaluate data mining courses at university level have been published so far, most of them in the computer science education literature.

An approach to teaching data mining at undergraduate level is described in (Lopez and Ludwig 2001). Their course was using the *Weka* open source data mining toolbox and accompanying text book (Witten and Frank 2000). Students initially had to research a data mining topic of their choice and present their findings, followed by lectures covering the major data mining topics. Practical exercises were conducted using *Weka*, which proved to be a useful learning tool. In the final part of the course, students had to perform a data mining group project

Copyright ©2007, Australian Computer Society, Inc. This paper appeared at the Sixth Australasian Data Mining Conference (AusDM 2007), Gold Coast, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 70. Peter Christen, Paul Kennedy, Jiuyong Li, Inna Kolyshkina and Graham Williams, Ed. Reproduction for academic, not-for profit purposes permitted provided this text is included.



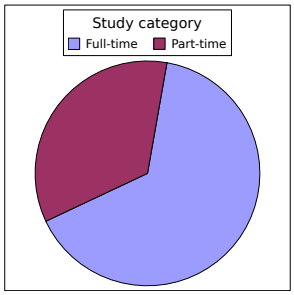
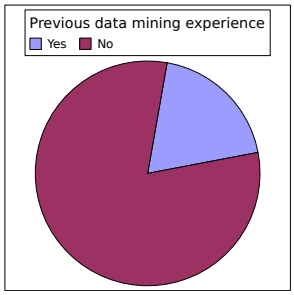


Figure 1: Student degrees and categories.



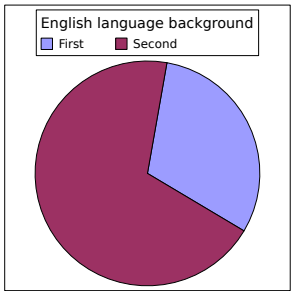


Figure 2: Student's previous data mining experience and English language backgrounds.

using a data set with around 250,000 records provided by the U.S. Forest Service. The authors report that the course was a success, however, they do not discuss what prior knowledge was expected from the students, nor do they present an empirical evaluation of their course.

A rather different approach is presented in (Musicant 2006). This data mining course for computer science undergraduate students is based mainly on selected research papers. The author builds the assignments for the course on these papers, asking students to read them, post questions onto a discussion forum, and then implement the algorithms described in the papers. In a final project students were free to either implement an advanced data mining algorithm of their choice, or conduct a data mining study on data sets of their choosing. The author then remarks that his course is unique as it exposes students to reading research papers and requires their critical thinking by coming up with relevant questions about these papers. He however concludes that marking and grading of the implementation and data mining studies is the most challenging part of his course.

A very recent report on a data mining course aimed at both undergraduate and graduate computer science and non-computer science students is presented in (Saquer 2007). The emphasis of this course was to provide an understanding of the basic (and some advanced) data mining algorithms, with a focus on classification, clustering and association rules. Each of these three topics received about four weeks out of a sixteen week course, which started with an introduction and finished with a group project where student either had to implement an advanced algorithm or learn and present about a data mining topic not covered in the lectures. Two assignments were conducted where students had to select a data set of their choice, use two algorithms to carry out a data mining project, and write a report of their findings. There is no thorough course evaluation provided in this paper; however, the author reports positive student feedback.

3 Course Student Backgrounds

Of the initial 27 students enrolled in the data mining course discussed here, one dropped out throughout the semester, and only one of the remaining 26 students was female. The details of student degrees and study categories are shown in Figure 1, while Figure 2 provides a breakdown in student's previous experience in data mining and their language background (if they were native English speakers or not).

Eighteen students were enrolled in a course-work masters program, while three were fourth-year computer science honours students, and five of the students were external participants that only enrolled for this course (non-award enrolment) and came from various private and public sector organisations (mainly from government agencies). Around 60% of the students were studying full-time.

At the beginning of the course, five of the 26 students indicated they had previous data mining experience (this included three of the five external participants), and three had attended a data mining course before (including one of the five external participants). Almost 70% of all students had English as a second language, with several students having arrived in Australia just this year for their one-year masters course-work studies.

4 Course Structure

The course consisted of nineteen one-hour lectures as summarised in Table 1, with the first five modules presented before the semester break and the rest afterwards. Modules 4, 5, 6, 7 and 9 were mostly based on corresponding chapters from the text book *Data Mining: Concepts and Techniques* (Han and Kamber 2006) which was used in the course, while the other modules were a mix of text book based material and additional material developed by the author. The *End-to-end data mining* lecture was given by an industry based data miner with many years of practical and teaching experience.

Four practical one-hour laboratory sessions were conducted using the open source tool *Rattle* (Williams 2007)[1], which is a graphical user interface built on top of the *R* statistical programming language[2]. *Rattle* provides a logical user interface to many data mining algorithms implemented in *R*, and includes techniques for data exploration and transformation, clustering, association rule mining, various classification techniques, and a variety of evaluation methods. It also allows direct user access to the underlying *R* console and thus facilitates further exploration of data mining algorithms as well as statistical concepts.

In the first practical laboratory session *Rattle* was introduced to the students and a small data set was explored. The second session covered association rule mining, and the students had to conduct various experiments using a publicly available data set sourced from the UCI machine learning repository (Newman et al. 1998). In the third and fourth laboratory sessions students were asked to work with decision tree and support vector machine classifiers, and compare their performance using various evaluation methods, such as ROC, risk and precision-recall curves. *Rattle* was also used for the assignments as discussed in the following section.

Laboratory sessions were held around every two weeks, with tutorials in the intermediary weeks. For each tutorial, students were asked to read two data

[1] http://rattle.togaware.com
[2] http://www.r-project.org

Module	Topic	Hours
1	Course introduction and data mining overview	1
2	Data mining process and data issues in data mining	2
3	Data pre-processing and data integration and linkage	2
4	Mining frequent patterns and associations	2
5	Cluster analysis	2
6	Classification and prediction	4
7	Mining time series and data streams	1
8	Privacy-preserving data mining	1
9	Web and text data mining	2
10	End-to-end data mining, data mining trends and social impacts	2

Table 1: Course lectures overview.

Tutorial	Papers discussed
1	*Data cleaning: Problems and current approaches* (Rahm and Do 2000) *Methods for evaluating and creating data quality* (Winkler 2004)
2	*Fast algorithms for mining association rules* (Agrawal and Srikant 1994) *Selecting the right interestingness measure for association patterns* (Tan et al. 2002)
3	*On comparing classifiers: Pitfalls to avoid and a recommended approach* (Salzberg 1997) *Classifier technology and the illusion of progress* (Hand 2006)
4	*State-of-the-art in privacy preserving data mining* (Verykios et al. 2004) *Names: A new frontier in text mining* (Patman and Thompson 2003)

Table 2: Tutorial papers.

mining papers, selected by the author and listed in Table 2, which were then discussed in the tutorial sessions. These tutorials exposed students to a broad range of topics and were aimed at deepening and complementing the topics covered in the lectures and laboratory sessions, as well as for students to critically read and analyse research papers, and to encourage discussion on these papers. Reading these tutorial papers was also aimed at preparing students for their presentation of a research paper in the final semester week, as discussed in the following section.

All course material was made available on a Web site, with lecture slides normally being uploaded two days before a lecture. Besides lectures, tutorials and laboratory sessions, an electronic discussion board (forum) was set up to allow dissemination of information from the lecturer to students, and to enable students to post questions and discuss them among themselves. Students were encouraged to post questions into this forum rather than e-mail them to the lecturer, with the aim to initiate discussions between students. This turned out to be less successful than expected: most posted messages were student questions, that were followed by answers and clarifications by the lecturer, but there was almost no discussion among students. However, most students reported in the end-of-semester questionnaire to have used the forum regularly as readers, and they commented that it was a useful tool for getting information.

5 Course Assessment

The assessment for the course consisted of two assignments, each worth 15% of the final course mark; a paper presentation and report in the last semester week (worth 20%); and a final written examination which was worth 50% of the final course mark.

The first assignment consisted of two parts, the first being a two-page essay on data issues related to a university data warehouse (such as what data a university is collecting, how it would design a data warehouse, what kind of data mining it would be interested in, and what the data mining challenges would be in such an environment). For the second part, students could choose to either implement, test and evaluate a simple clustering algorithm (such as k-means), or conduct a clustering project using the *Rattle* tool on a publicly available data set of their choice. For both options they had to write a four-page report detailing their findings. Eight students choose to implement a clustering algorithm, all of them masters students, while all others conducted the clustering project. Somewhat surprisingly, all three computer science honours students selected the clustering project rather than implementing an algorithm.

For the second assignment, students had to conduct association rule mining on a small data set made of seven transactions; had to calculate different classifier accuracy measures (such as precision, recall and specificity) on a confusion matrix that was based on their seven-digit university identifier (for example, for an identifier 1234567, the number of true positives were the first four digits 1234, the number of false positives the last three digits 567, the number of true negatives were the first four digits of the reversed identifier 7654, and the number of false negatives the last three digits of the reversed identifier 321); and they had to conduct a classification project using *Rattle*, comparing three different classification techniques (decision trees, support vector machines, and a third classifier of their choice) on a publicly available data set of their choice sourced from the UCI machine learning repository (Newman et al. 1998). For the last part they had to write a report that had to include a ten-line executive summary, details of the data exploration and transformation steps they have conducted, a description of the classifier approach they have taken, details of the results they have achieved (including confusion matrices and ROC curves for all three classifiers), a critical summary of their project, and a reflection of what they have learnt.

The final student presentation was originally planned to consist of a 15-minutes talk by each student, but due to the larger enrolment number this had to be changed to 'lightning' talks of five minutes per student (still resulting in three one-hour sessions of presentations). Students were able to select a data mining research paper of their choice (with the only limitation that it had to be published from the year 2000 onwards), and then had to write a report addressing the techniques described in the paper, data sets used and experiments conducted, measurements employed to assess the quality of complexity of the techniques described, and finally a critical assessment of the paper (detailing, for example, limited experimental studies on only small or only synthetic data sets, claims written by the authors that were not supported by theory or experiments, etc.). They then had to summarise their report onto four slides and give a short presentation. They were able to receive up to ten marks for their report, and up to five marks each for their slides and their oral presentation.

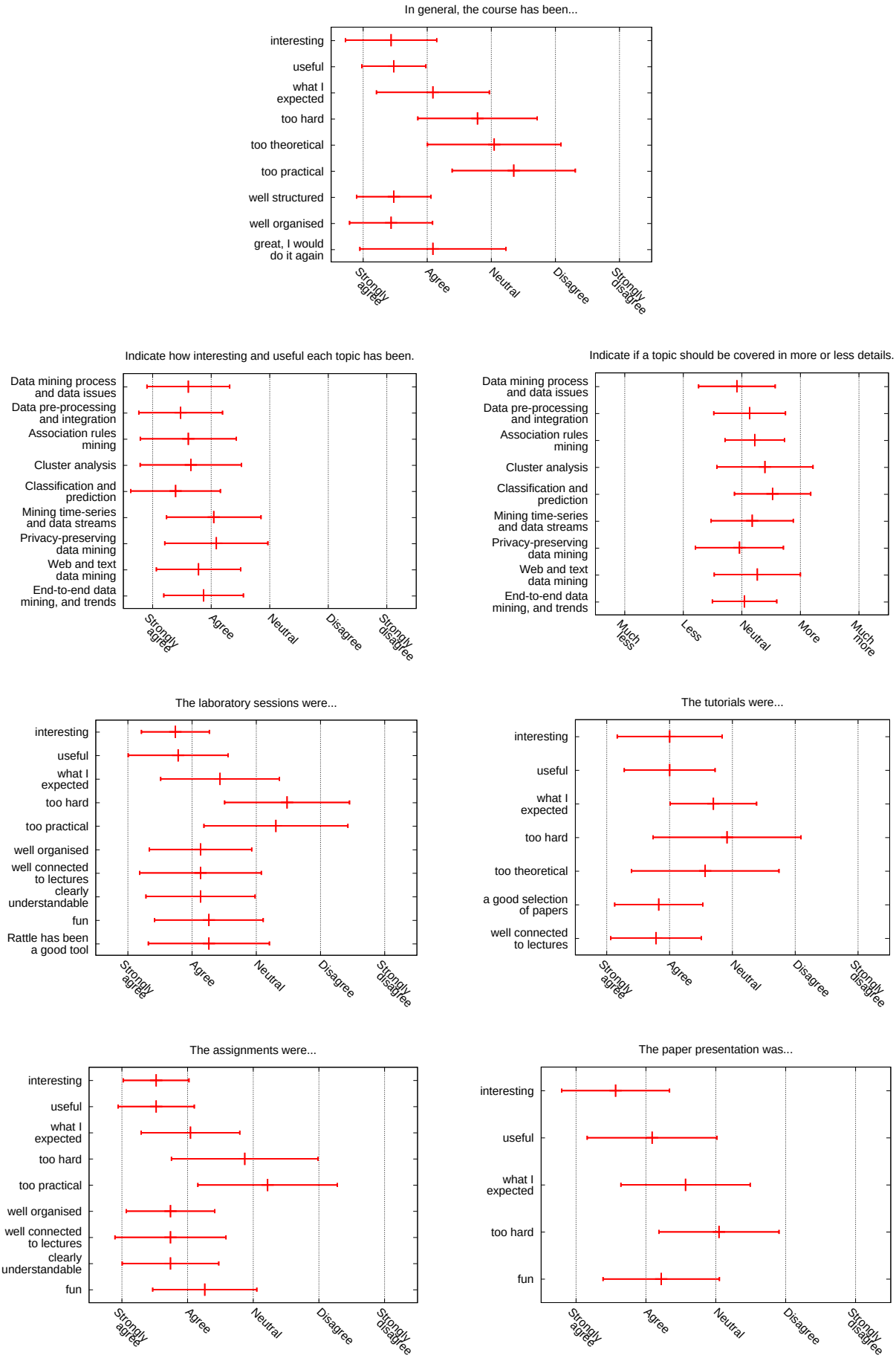


Figure 3: Course evaluation results from all students (mean and standard deviation).

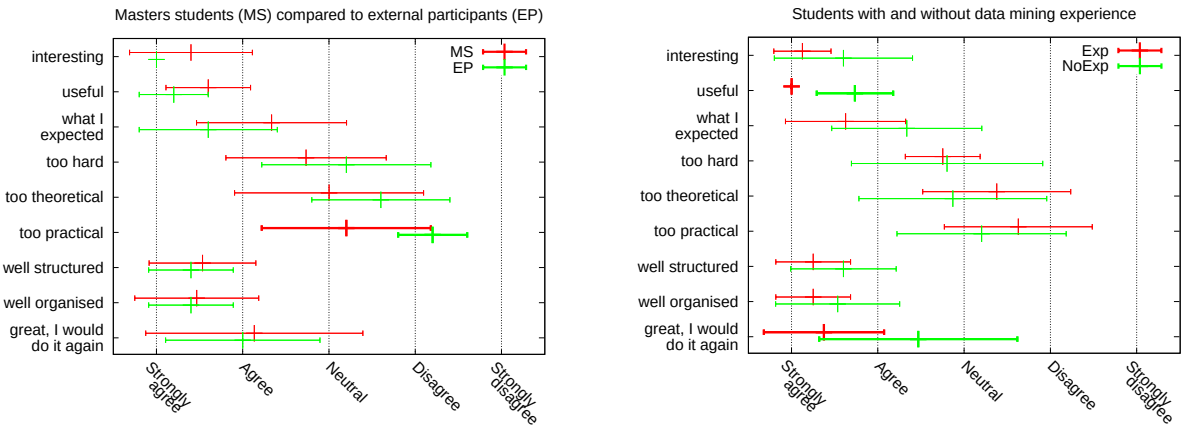


Figure 4: In general, the course has been... (Statistically significant differences ($p < 0.05$) are highlighted bold in this and all following figures.)

Half of the total course mark was based on a final three-hour written examination. While this was not an open book examination, the students were allowed to take one sheet of A4 paper with written notes into the examination. The emphasis of the questions, which covered the whole course material, was on explaining concepts rather than formulas, definitions or implementation details of algorithms.

Overall, the students put a lot of effort into their assignments, and especially into their presentations, which resulted in excellent written project reports and oral presentations. The high quality of the submitted material is reflected in good average student marks, as will be discussed in Section 6.1 below.

6 Course Evaluation and Discussion

At the end of the semester a four-page questionnaire, designed by the author, was handed out to all 26 students. Given that this was a new course, run for the first time in 2007, the main objective of this questionnaire was to get feedback about the course that would allow improvements in coming years. The questionnaire consisted of seven tables where students had to rate the different aspects of the course, plus eleven free-format questions allowing students to provide additional comments. The main results from the seven questionnaire tables are shown in Figures 3 to 10, and they will be discussed in detail in Section 6.2 below. These figures display mean and standard deviation values. First, in the following section, a short overview of the student marks from the various course assessments is provided.

6.1 Distribution of Student Marks

The quality of the submitted reports has been mainly good for all four assessments of this course. The average mark for the first assignment was 10.8 (out of 15), with a standard deviation of 1.9 and marks ranging from 5.5 to 13.5. For the second assignment, the marks were higher, with a mean of 13.0 (out of 15), a standard deviation of 1.6 and marks from 9.5 to 14.5. The student presentation marks had a mean of 16.5 (out of 20) with a standard deviation of 1.6, and marks ranging from 13.5 to 19.5. The average final exam mark (out of 50) was 37.7, with a standard deviation of 9.6 and marks ranging from 21 to 48.5. All students passed the course, with ten students achieving a high-distinction course mark (80 or above out of 100) and eleven a distinction (between 70 and 79). Only five students had a course mark below 70.

6.2 Questionnaire Results

Of the 26 students enrolled, 23 returned the questionnaire, including all three computer science honours students and all five external participants. The overall results of all returned questionnaires are shown in Figure 3. As can be seen, general feedback was very positive, with most students agreeing that the course had been useful and interesting, and that it had an appropriate mix of theory and practice. Most of the topics covered in the course were well received, with more detailed coverage desired mainly for the core topics of clustering, classification and prediction. The laboratory sessions could have been made slightly harder and more practical, while some of the papers discussed in the tutorials were felt to be too theoretical. In hindsight, it would have been advantageous to have provided additional material, containing mainly statistical and mathematical background information, together with the tutorial papers.

In order to be able to discuss the differences in feedback between graduate (masters) students (15 responses) and external participants (5 responses), as well as to differentiate between those who indicated they had previous data mining expertise (8 responses) and those who did not (15 responses), Figures 4 to 10 show two graphs each with the corresponding student sub-groups. Note that in the analysis of masters students and external participants (left side in the figures) the results of the three honours students are not included. However, they are included in the graphs showing previous data mining expertise or not (right side in the figures). Statistically significant differences in the results, as measured with a $p < 0.05$ confidence using the two-tailed Mann-Whitney U test (Sheskin 2004), are shown in bold.

6.2.1 General Course Impressions

Figure 4 shows the results of the question *'In general, the course has been...'*. The main, statistically significant, difference between masters students and external participants was that the external participants would have liked the course to be more practical. Other differences between masters students and external participants were that the course was closer to what the external participants expected; that masters students found the courses harder; and that external participants also would have liked the course to contain more theory. A similar distinction can be seen between those students who had previous data mining experience and those who did not. Significant

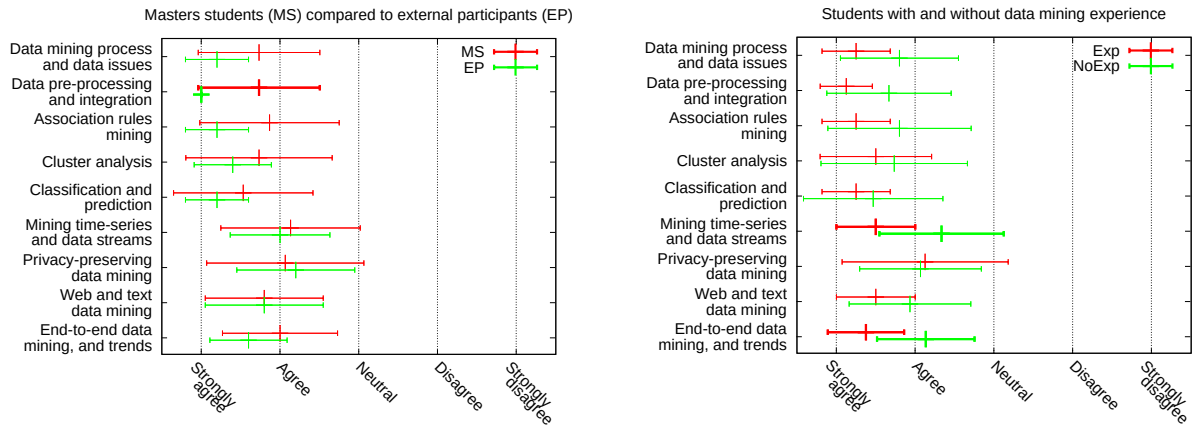


Figure 5: Indicate how interesting and useful each topic has been.

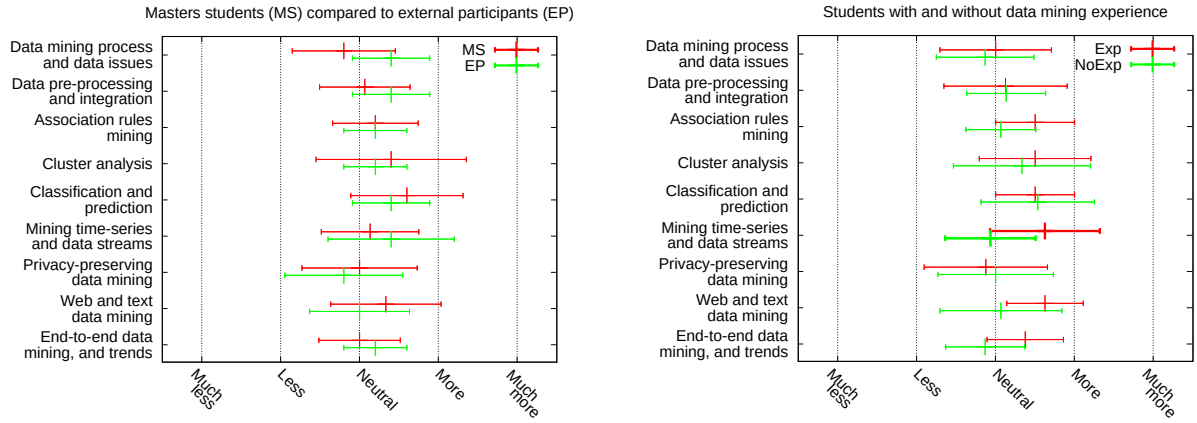


Figure 6: Indicate if a topic should be covered in more or less detail.

differences were that those with previous experience found the course to be more useful (they all strongly agreed that the course was useful) and that they more likely would do such a course again. Those with previous experience also found the course more interesting, they agreed more in that the course was what they expected, and they would have liked the course to be both more theoretical and more practical.

Some students commented in their questionnaire that they appreciated the clear communication by the lecturer, which included schedules, material put online well in time, as well as clear formulation of what was expected in assignments. Many students also commented that they liked the text book used (Han and Kamber 2006), as it allowed them to read more about the concepts that were only covered briefly in the lectures.

6.2.2 Interestingness and Usefulness of Topics

The results on interestingness and usefulness of the topics covered in the course are shown in Figure 5. A statistically significant difference was that the external participants agreed stronger that the data preprocessing and integration topics were more interesting and useful compared to the masters students. This is likely to be because the external participants had experience with real world data issues beforehand, and are thus more aware of the importance of these topics. Other significant differences were that the students with previous data mining experience rated the time-series and data streams topic, as well as the *End-to-end data mining* and social impacts lectures, more interesting and useful than those without previous experience. With the exception of privacy-preserving data mining and Web and text mining, the external participants rated all other topics to be more interesting and useful than the masters students did. Those with previous data mining experience also rated all topics, except privacy-preserving data mining, more interesting and useful than those without previous experience.

6.2.3 Coverage of Topics

As Figure 6 shows, the only statistically significant difference with regard to the coverage of topics was that the students with previous data mining experience wanted to hear more about data stream and time-series data mining. Other differences included that masters students mainly wanted to have more coverage of clustering, classification and prediction, as well as text and Web mining. The external participants, on the other hand, would have preferred to also learn more about the data related first two topics, and classification and prediction. Students without previous data mining experience would have liked more detailed coverage of the classification and prediction topics, while those with previous experience would have preferred increased coverage of almost all topics, with the exception of privacy-preserving data mining and the initial data mining process and data issues topics. They were very likely already famil-

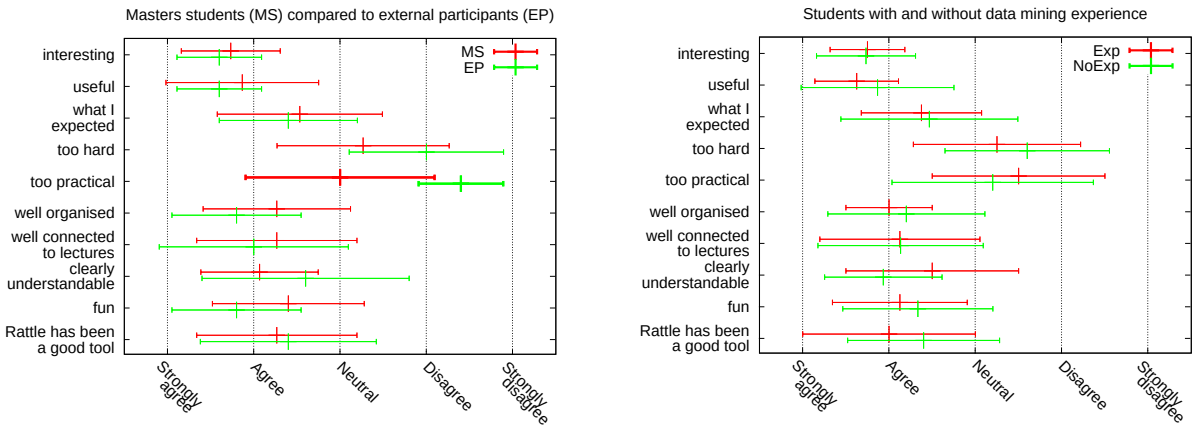


Figure 7: The laboratory sessions were...

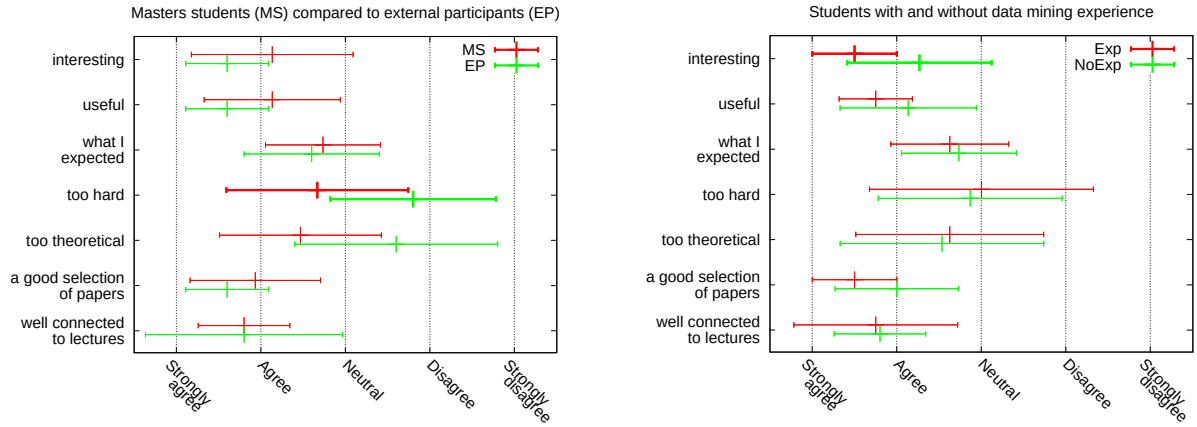


Figure 8: The tutorials were...

iar with the introductory issues discussed in this first topic. Some students commented in the questionnaire that the *End-to-end data mining* lecture was one of the best parts of the course.

Several students wrote in the questionnaire that they would have liked topics to be covered in more detail rather than just having overviews of many concepts (such as introductions to many recently developed clustering techniques). Others would have preferred to hear about the core data mining techniques in more depth. One way to allow more detailed coverage in the future will be to increase the number of lectures. Mentioned several times in the questionnaire was the wish to include more practical examples into the lectures. Some students would also have liked to either have more basic explanations of the mathematical and statistical concepts, or provision of resources that contained such background material.

6.2.4 Practical Laboratory Sessions

The results about laboratory sessions in Figure 7 show a clear distinction between masters students and external participants, in that the latter group would have preferred the laboratories to be both more practical (a statistically significant difference) and harder. On the other hand, the external participants found the laboratory sessions to be less understandable, which was likely due to a different computing environment used at the university compared to their workplaces, as well as that the masters students were more familiar with the practice of computer science laboratory sessions. There were no significant differences between those students with previous data mining experience and those without. However, those with previous experience found the laboratory sessions to be less understandable than those without experience. This might be due to the fact that they have been using a different data mining software previously, and were unfamiliar with the *Rattle* tool used in this course.

Several students criticised the use of *Rattle* as the only data mining tool presented in the course. They would have preferred to learn more about other available tools, or even use different tools in the laboratory sessions and assignments. Other students commented that they would have liked more documentation on the functionality of the software, including detailed explanations of how parameter settings influence algorithms. Some students also wrote that they would have liked to have practical demonstrations of data mining tools in the lectures. Another criticism was that the initial installation of the *Rattle* tool had some minor deficiencies at the beginning of the semester and was not very stable in several instances (a potential risk when using any open source tool that is still under development).

6.2.5 Research Paper Tutorials

The main statistically significant differences in feedback on the tutorials, as shown in Figure 8, are that the masters students found these tutorials much harder than the external participants, and that the

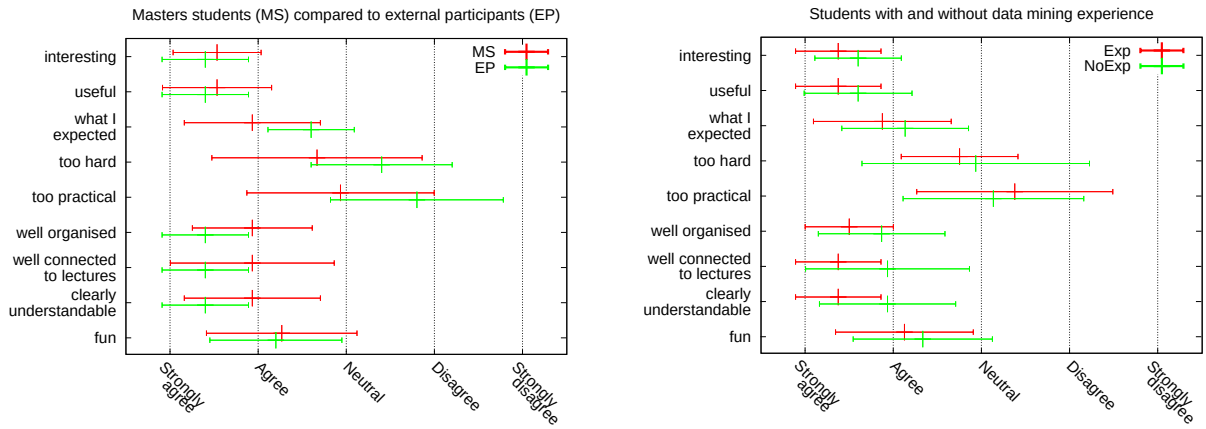


Figure 9: The assignments were...

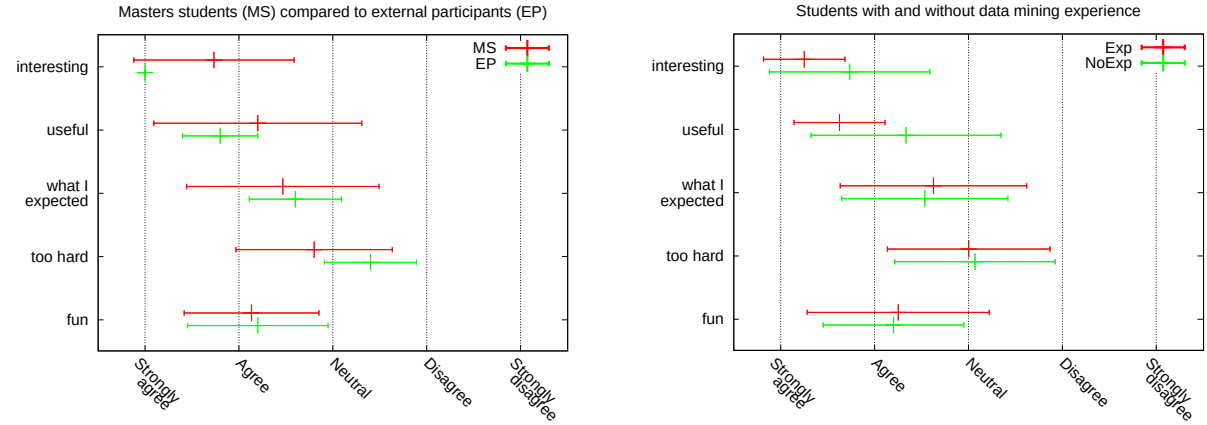


Figure 10: The paper presentation was...

students with previous data mining experience found the tutorials to be more interesting than those without previous experiences. The external participants also reported that reading these papers was more useful and interesting than the masters students did, and they also found these papers to be less theoretical to read than the masters students. On the other hand, those with and without previous data mining experience both reported similar agreements with regard to how hard and how theoretical the papers were. However, those with previous experience agreed stronger that the papers selected were appropriate, and they also found the tutorials to be more useful.

One explanation for these differences is that, due to students other commitments and room limitations, two tutorial groups were run, the first with ten students (all of them masters students) while the second group with sixteen students included all external participants. This resulted in rather different discussions of the tutorial topics covered. For future offerings of such tutorials based on paper readings, either a more balanced distribution of students should be aimed for, or even better only one single tutorial group that would include all students. Comments provided in the questionnaires included that students would have liked the tutorials to be more connected to both lectures and the text book, to only include papers with less mathematical and statistical content, as well as that specific questions were provided by the lecturer for each paper to make the reading more specific. Some students would also have preferred more practical laboratory sessions rather than tutorials.

6.2.6 Assignments

The feedback on the two assignments, shown in Figure 9, is mainly positive. There are no statistically significant differences between the two pairs of student sub-groups. The main differences between masters students and external participants were that the former found the assignments harder but practical enough, while the external participants would have preferred to have assignments of a more practical form. The external participants were also more agreeing that the assignments were well organised and understandable, as well as better connected to the lectures than the masters students.

On the other hand, masters students were more agreeing that the assignments were what they expected. This is likely due to their experience of assignments in other courses they attend, while the external participants only attended this one course and thus had less experience in what to expect. Those with previous data mining experience found the assignments less hard than those without previous experience, but better organised and understandable, and better connected to the lectures.

Student comments on the assignments included that they would have liked more freedom in the use of data sets (for example, one external participant would have liked to use a data set from his workplace), with less emphasis on writing reports and more practical programming in the assignments.

6.2.7 Research Paper Presentation

The results for the paper presentation in Figure 10 are quite different for masters students and external participants, although none of these differences are statistically significant. Masters students found the work involved and giving a presentation harder but also less useful and less interesting. A comment from one external participant indicates that the large number of 26 presentations provided him with an exposure to a wide range of topics, including several papers in his area of interest. Similar differences in interestingness and usefulness can also be seen between those with and without previous data mining experience (again not statistically significant).

Some students critically commented in the questionnaire that the five minutes time given for a presentation was too short, and having the second assignment and student presentation both in the last semester week was a very high workload. Others would have preferred the presentations to be linked better with the tutorial sessions, as well as to have student presentations spread throughout the semester rather than having all of them in the last week.

7 Conclusions

This paper has discussed the student population, course structure and assessment of a new graduate level data mining course taught at a major Australian university in 2007, and presented an empirical evaluation of student perception of the course. Overall, the feedback was positive, with students reporting that the course has been interesting, useful, and that it contained about the right levels of theory and practice. They also reported to have liked the assessment consisting of two project based assignments and one presentation of a data mining research paper.

The results of the empirical evaluation indicate that it is possible to run a data mining course with both masters students and participants from industry and government organisations, even though these two groups likely have very different backgrounds of prior knowledge and experiences, and also different expectations in such a course. External participants, some with considerable expertise in data management and processing, were especially active during tutorial discussions of research papers, and contributed with their practical knowledge and experience. This was enriching the course experience for the students that so far did not have industry experience.

For future offerings of the data mining course described in this paper, there are several issues that can be improved. First, increasing the number of lectures will allow coverage of topics in more details. Second, having had two tutorial groups, one with only masters students, was clearly not optimal. A better arrangement would have been one single tutorial group only, to facilitate discussions and exchange of experiences and ideas between external participants and graduate students. Additionally, in order to make students think more actively and critically about the tutorial papers, it will be useful to ask them to provide written questions about these papers before the tutorial sessions, similar to (Musicant 2006). Third, allowing more choice in both selection of data sets and data mining tools for their assignments would allow students to better explore areas of their interest.

8 Acknowledgements

The author would like to thank the students of COMP8400 (semester 1, 2007) for providing their feedback on this course, and to Tom Gedeon and Paul Thomas for providing valuable feedback on a draft version of this paper and for proof-reading it.

References

Agrawal, R. & Srikant, R. (1994), Fast algorithms for mining association rules in large databases, *in* 'International Conference on Very Large Data Bases', Santiago de Chile, Chile, pp. 487–499.

Han, J. & Kamber, M. (2006), *Data Mining: Concepts and Techniques*, 2nd edition, Morgan Kaufmann.

Hand, D.J. (2006), 'Classifier technology and the illusion of progress', *Statistical Science*, vol. 21, no. 1, pp. 1–14.

Lopez, D. & Ludwig, L. (2001), Data mining at the undergraduate level, *in* 'Midwest Instruction and Computing Symposium', Cedar Falls, Iowa.

Musicant, D.R. (2006), A data mining course for computer science: Primary sources and implementations, *in* 'SIGCSE-06: Proceedings of the 37th SIGCSE technical symposium on computer science education', ACM Press, Houston, Texas, pp. 538–542.

Newman, D.J., Hettich, S., Blake, C.L. & Merz, C.J. (1998), UCI Repository of machine learning databases, `http://www.ics.uci.edu/~mlearn/MLRepository.html`, University of California, Department of Information and Computer Science, Irvine, California.

Patman, F. & Thompson, P. (2003), Names: A new frontier in text mining, *in* 'First NSF/NIJ Symposium (ISI-2003)', Tucson, AZ, Springer LNCS 2665, pp. 27–38.

Rahm, E. & Do, H.H. (2000), 'Data Cleaning: Problems and Current Approaches', *IEEE Data Engineering Bulletin*, vol. 23, no. 4, pp. 3–13.

Salzberg, S.L. (1997), 'On comparing classifiers: Pitfalls to avoid and a recommended approach', *Data Mining and Knowledge Discovery*, Springer, vol. 1, no. 3, pp. 317–328.

Saquer, J. (2007), 'A data mining course for computer science and non-computer science students', *J. Comput. Small Coll.* , vol. 22, no. 4, pp. 109–114, Consortium for Computing Sciences in Colleges.

Sheskin, D.J. (2004), *The Handbook of Parametric and Nonparametric Statistical Procedures*, 3rd edition, Chapman & Hall/CRC.

Tan, P.N., Kumar, V. & Srivastava, J. (2002), Selecting the right interestingness measure for association patterns, *in* 'ACM SIGKDD international conference on knowledge discovery and data mining', Edmonton, Canada, pp. 32–41.

Verykios, V.S., Bertino, E., Fovino, I.N., Provenza, L.P., Saygin, Y., & Theodoridis, Y. (2004), 'State-of-the-art in privacy preserving data mining', *SIGMOD Record*, vol. 33, no. 1, pp. 50-57.

Williams, G.J. (2007), 'Data Mining with Rattle and R', Togaware, Canberra, `http://datamining.togaware.com/survivor/`.

Winkler, W.E. (2004), 'Methods for evaluating and creating data quality', *Information Systems*, Elsevier, vol. 29, no. 7, pp. 531–550.

Witten, I.H. & Frank, E. (2000), *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann.