

Reflection on Development and Delivery of a Data Mining Unit

Bozena Stewart

School of Computing and Mathematics
University of Western Sydney
Locked Bag 1797 Penrith South DC NSW 1797

b.stewart@uws.edu.au

Abstract

Educators developing data mining courses face a difficult task of designing curricula that are adaptable, have solid foundations, and are tailored to students from different academic fields. This task could be facilitated by debating and sharing the ideas and experiences gained from the practice of data mining as well as from teaching data mining. The shared body of knowledge would be a valuable resource which would help educators design better data mining curricula. The aim of this paper is to make a contribution to such a debate. The paper presents a reflection and evaluation of the author's experience with developing and delivering a postgraduate unit Knowledge Discovery and Data Mining.

Keywords: curriculum, data mining, education, teaching

1 Introduction

Advances in computer technologies during the past decade made it possible to generate and collect vast amounts of data in many areas of human activities. The data contains hidden information that needs to be extracted and analysed to provide rules, patterns and models suitable for use in decision making.

The field of Knowledge Discovery and Data Mining (KDDM) has emerged in response to the practical need to analyse huge quantities of data collected in commerce and industry. KDDM has evolved as an interdisciplinary field at the intersection of machine learning, statistics, artificial intelligence, and database systems.

To fully utilise the power of data mining, industry and commerce need a steady supply of highly trained data mining analysts. The universities in Australia and around the world have responded to this need by creating specialised units or complete courses devoted to knowledge discovery and data mining.

In this paper the term "unit" is used to represent a component of a course. An alternative commonly used name is "subject". The term "course" is used here to

represent an academic program leading to a degree or a diploma award. A course is a collection of units.

So far there have been few guidelines published in the literature to help unit and course designers develop data mining curricula. In most cases, data mining courses and units have been developed in ad hoc manner, typically by individual academics according to their own experience and research interests. Such curricula are often biased towards the designer's favourite topics and may not adequately cover topics required for data mining analysts employed in industry or commerce.

Recently, in response to the developments in the field the ACM SIGKDD Executive Committee setup the ACM SIGKDD Curriculum Committee to design a sample curriculum for data mining. The committee released the first draft proposal in April 2006 (SIGKDD, 2006). The proposed curriculum contains a comprehensive set of topics and guidelines which will undoubtedly become the basis of many data mining courses in the future. The curriculum is a work in progress which still needs to include sample units (subjects) to provide educators with the guidance on how to structure and package topics for students from different disciplines.

The field of data mining is expanding at a fast pace as new data mining algorithms and methodologies are invented and applied to new domains. The dynamic nature of data mining demands a curriculum that can adapt to changing needs. Educators face a difficult task of developing units and courses that are dynamic, have solid foundations, and that are tailored to students from different academic disciplines. This task could be facilitated by debating and sharing the ideas and experiences gained from the practice of data mining as well as from teaching data mining. The shared body of knowledge would be a valuable resource which would help educators design better data mining curricula. The aim of this paper is to make a contribution to such a debate. The paper describes the author's experience with teaching a postgraduate coursework masters unit *Knowledge Discovery and Data Mining*.

The remainder of this paper is organised as follows. Section 2 surveys related work, Sections 3 to 5 discuss the background of the data mining unit, the unit objectives, the unit content, and the unit assessment. The student feedback on the unit is analysed in Section 6 and the author's reflection on the unit is given in Section 7, followed by a summary of the paper in Section 8.

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.

2 Related Work

There have been relatively few papers published in the literature describing the design and evaluation of data mining units. In recent years there have been several articles published in the educational literature reporting on experience with designing units for Computer Science majors. For example, Musicant (2006) designed a unit focused on "computer science" aspects of data mining. He used specific research papers and assignments that allowed students to implement data mining algorithms themselves. This approach allows students to identify more closely with data mining researchers and challenges the students to understand the intricacies of data mining algorithms.

Chawla (2005) reports on experience in offering a data mining unit to upper level undergraduate and postgraduate students in computer science and engineering. The unit was based on Witten and Frank (1999) text and utilised the Weka data mining tool and Matlab software. The unit incorporated readings of research papers and a conference-style research project.

Saquer (2007) discusses the design and evaluation of a data mining unit taught successfully to computer science and non-computer science majors. The unit was designed for both undergraduate and graduate students. To provide for different student backgrounds, the unit allowed non-implementation as well implementation projects. Graduate students were required to read at least 3 research papers and write a report on them.

There are some similarities between the KDDM unit described in this paper and the data mining units mentioned above. The KDDM unit and Chawla (2005) and Saquer (2007) had a similar aim of providing grounding in theoretical and practical aspects of data mining, all utilised packages for practical work, and all contained a project. The KDDM unit was similar to Saquer (2007) in that both units were targeted to students with different backgrounds. KDDM's project, however, did not involve implementation of data mining algorithms and did not include a conference-style research paper.

3 Unit Background

The data mining unit described in this paper was first created by the author in 2003 as a special topic for the unit *Emerging Issues in Computing*. The term "special topic" is used here as a synonym for content. The purpose of the *Emerging Issues in Computing* unit was to introduce students to new areas of computing. The content of the unit could vary from year to year, depending on which emerging area was being taught. In 2003 the whole unit was devoted to data mining. In the following year, the author created a new unit named *Knowledge Discovery and Data Mining (KDDM)*. This unit was designed as an elective unit for the Software Engineering major in the coursework masters degree Master of Computing. Masters students from other majors also were allowed to enrol, provided that they satisfied the prerequisite requirements. The content and objectives

of the new unit were similar to the previous unit but the teaching materials were updated and streamlined to better suit the new unit.

The aim of the KDDM unit was to provide students with a broad knowledge of the basic concepts and techniques used in data mining.

In 2005, unfortunately, due to the rationalisation of computing courses and units, the software engineering major in Master of Computing degree was discontinued and most of its units were closed, including the KDDM unit. Hence there were no further offerings of the KDDM unit beyond 2004.

3.1 Unit Content

The content of the unit covered the core data mining topics including data preparation, association rule mining, classification and clustering, and gave a brief overview of several advanced topics including text mining, Web mining, spatial data mining, and temporal mining. More details on the unit content are provided in Section 4.

3.2 Unit Delivery

The unit was delivered in the traditional face-to-face format consisting of a two-hour lecture followed by one-hour tutorial per week for 13 weeks. Tutorials were used to solve theoretical exercises supplementing the topics discussed in lectures. There were no formal laboratory classes but the students were expected to work on practical tasks in their own time. They could do their practical work in the computing laboratory equipped with the data mining tool Clementine from SPSS or they could do it at home using any suitable open source software such as WEKA.

3.3 Teaching Resources

The textbook used in the unit was Han & Kamber (2001). Another recommended text was Dunham (2003). The unit teaching materials were prepared by the author and were available to students online on the WebCT site. To motivate the students to explore data mining topics beyond the scope of the unit, the author also provided references to relevant journal and conference papers found in the ACM and IEEE digital libraries.

3.4 The Students

The students were full-fee paying students, the majority of whom were from overseas. There were 23 students at the start of the semester and 21 students at the end. The prior educational requirement for Master of Computing was an undergraduate degree in Computer Science, Information Technology, Information Systems, or similar. The assumed knowledge for the KDDM unit was a basic background in statistics, database systems and computer programming.

4 Unit Objectives and Unit Content

The unit objectives were to provide students with the knowledge and skills that would enable them to:

1. Understand the knowledge discovery process
2. Choose an appropriate strategy for data preparation
3. Select appropriate data mining algorithms for different applications
4. Compare and contrast each of the following data mining techniques: association rules, decision trees, neural networks, Bayesian networks, the nearest neighbour algorithm, support vector machines, clustering algorithms, and genetic algorithms
5. Apply data mining algorithms for association rule mining, classification and clustering to solving practical problems
6. Use data mining tools to solve a variety of data mining problems
7. Describe the OLAP operations and the design of a data warehouse
8. Explain the issues related to mining of unstructured and semi-structured data: text data, web data, time series data, and spatial data
9. Pursue higher research studies in data mining

weekly schedule of topics is listed below.

1. An overview of data mining
2. Data preparation
3. Association rule mining
4. Decision trees
5. Neural networks
6. Bayesian networks
7. Nearest neighbour; Support vector machines; Regression
8. Clustering
9. Genetic algorithms
10. Data warehousing and OLAP
11. Text mining; Web mining
12. Spatial mining
13. Temporal mining

To see how the unit content matched the unit objectives, the relationships between the unit topics and the unit objectives are shown in Table 1. The rows in the table represent the weekly topics and the columns represent the unit objectives. The plus (+) symbol in a given row and column indicates that the topic in that row contributed to achieving the objective in the corresponding column.

Topics	Unit Objectives								
	1	2	3	4	5	6	7	8	9
1	+								+
2		+				+			+
3			+	+	+	+			+
4			+	+	+	+			+
5			+	+	+	+			+
6			+	+	+	+			+
7			+	+	+	+			+
8			+	+	+	+			+
9			+	+	+	+			+
10							+		+
11								+	+
12								+	+
13								+	+

Table 1: Unit topics versus unit objectives

The unit content was designed to satisfy the unit objectives above. Topics were chosen on the basis of the author's experience and literature survey of data mining textbooks. The emphasis of the content was on the major data mining concepts and algorithms. In addition, the basics of data warehousing and OLAP, and several advanced data mining topics were briefly introduced in the last four weeks of the semester. The main reason for including the topic on data warehousing and OLAP so late in the semester was that the unit did not use data sets stored in databases, and at that time the author did not consider it necessary to introduce this topic earlier. The

The distribution of the '+' symbols in Table 1 shows that all the unit objectives were covered by the content of the unit. For example, the objectives 1, 2 and 7 were discussed in one topic (lecture) each, while the objectives 3 to 6 were addressed by 7 or more topics. This highlights the fact that the emphasis of the unit was on the core data mining algorithms and their associated concepts.

The objective 9, "Pursue higher research studies in data mining", requires knowledge of all core data mining topics as well as knowledge of more advanced material. Hence it could be argued that all of the topics covered in the unit contributed in some way to the students' ability to undertake higher research studies in data mining.

5 Unit Assessment

The unit was assessed on the basis of two components: Continuous Assessment (60%) and Final Examination (40%). Continuous assessment consisted of two individual assignments, each worth 15%, and a group project worth 30% of the total mark. The final examination was an open book 3 hour examination with 10 minutes of reading time.

5.1 Assignments

In both assignments, the students were required to perform practical experiments and submit a written report describing the data mining approach taken and the results obtained. To perform practical experiments the students used data mining tools, either Clementine in the computing laboratory or open source software at home.

Assignment 1 was focused on the topics of data preparation, association rule mining and decision trees. The students were required to solve two data mining problems. The first problem was to mine association rules from transactional data from a retail supermarket,

obtained from the FIMI Repository. The task involved determining suitable levels of minimum support and confidence, finding association rules, determining which rules were interesting, and suggesting how the retail supermarket could apply the interesting rules to increase its sales. In the second problem the students were given a data set from a banking domain containing missing values, errors, and redundant attributes and they had to clean the data and then use the prepared data set to mine decision trees. The students were required to experiment with alternative approaches to data preparation, evaluate resulting decision trees using cross-validation, and use the decision trees to predict unknown labels in a given test set. They submitted a report describing the experiments, results obtained, and explaining the effects of data preparation on the resulting decision trees.

Assignment 2 was concerned with classification and prediction. Its main objective was to compare the predictive performance of several different classification algorithms on a moderately large real-world data set. The students could choose any three classification methods. They were given a training data set from an insurance domain containing an unbalanced class attribute and their first task was to create a balanced data set. Then they used the balanced data set to create 3 different classification models and used each model to predict class labels in a supplied test set. The students were required to evaluate the performance of each algorithm in terms of classification accuracy, precision and recall. In the report they were required to describe the experiments and the results obtained, and also explain how the results could be used to produce a mailing list for a direct marketing campaign.

5.2 Project

The project was a group project conducted in teams of 3 students. Its aim was to give the students practical experience with data mining of a large real-world data set. Each team selected a large data set from the UCI KDD Archive. The teams were free to use any appropriate data mining techniques and any tools available to analyse the data. In the middle of the semester, each team submitted a description of their chosen data set and the plans and timeline for the intended approach. At the end of the semester they submitted the final report presenting the results and describing the work in detail.

5.3 Unit Assessment versus Unit Objectives

To see how well the unit assessment corresponded to unit objectives, the relationships between the assessment tasks and the unit objectives are shown in Table 2. The rows of the table represent the assessment tasks and the columns represent the unit objectives.

The objective 7, concerned with basic concepts of OLAP and data warehousing, and the objective 8, concerned with advanced data mining, were assessed only by means of the final examination. They were introduced only briefly and represented a minor component of the unit content.

Assessment	Unit Objectives								
	1	2	3	4	5	6	7	8	9
Assignment 1	+	+	+	+	+	+			+
Assignment 2	+	+	+	+	+	+			+
Project	+	+	+	+	+	+			+
Final Exam	+	+	+	+			+	+	

Table 2: Assessment versus objectives

5.4 Student Performance

Have the unit objectives been achieved? This question can be partially answered by looking at the students' performance on the individual assessment tasks.

It can be seen from Table 3 that the students performed well on the specified assessment items. The individual assignments were in most cases of a good standard. Most of the group projects were of high standard, indicating that the students mastered the core concepts and techniques well. Students also performed relatively well on the final examination, with the average mark of 60%. The students' performance suggests that the unit met the unit objectives.

Assessment Task	Average Mark	Average %	Min Mark	Max Mark
Assign 1 (out of 15)	11.5	76.8	7.8	14.9
Assign 2 (out of 15)	13.0	86.8	9.9	15.0
Project (out of 30)	24.8	82.5	17.4	29.7
Final Exam (out of 40)	24.0	60.0	14.4	36.4
TOTAL (100)	73.3	73.3	50.1	92.5

Table 3: Assessment and student performance

6 Student Feedback

To obtain a feedback on the delivery of the unit, the author conducted the student evaluation survey (SEEQ) in the last week of the semester. Most of the questions on the SEEQ evaluation form were concerned with the teaching performance of the staff member and only a few questions related to more general aspects of the unit. The mean scores of responses on the more general questions are shown in Table 4 and the details of the distributions of student responses are given in Table 5 and Table 6. For comparison, the tables include the survey results for both data mining units, *Emerging Issues in Computing* (EIC) offered in 2003, and *Knowledge Discovery and Data*

Mining (KDDM) offered in 2004. There were 24 respondents out of 40 in 2003, and 10 respondents out of 21 in 2004.

The results in Table 4 show that the mean scores for the 2003 offering were considerably lower than the corresponding scores for 2004. These differences are also apparent in the distributions of the scores in Tables 5 and 6. However, a detailed comparison of the scores is difficult because of unequal number of respondents in the two evaluation surveys.

There were a number of reasons for the differences in the evaluation scores of the two units. One reason was that the student populations were quite different in 2003 and 2004. The *Emerging Issues in Computing* unit was a general elective not aimed at any particular major. As a result, the students had a wide range of backgrounds and some lacked adequate mathematical and computing knowledge. The *Knowledge Discovery and Data Mining* unit, on the other hand, was aimed at Software Engineering major students who had appropriate background for this unit and consequently faced few difficulties.

Another reason for the differences in the scores was that in 2003 the data mining unit was offered for the first time and the author had no prior experience with teaching a similar unit. In view of the student feedback, the author revised the syllabus and teaching materials for the second offering in 2004.

From the scores in Table 4 for 2004 survey and the bar charts of student responses in Tables 5 and 6 it can be concluded that:

- The respondents found the unit intellectually challenging and stimulating. They considered the content of the unit valuable and their interest in data mining was increased by doing this unit. They learned and understood the subject material in the class.
- They considered the methods of assessment to be appropriate, and the assignments and prescribed readings to be valuable and contributing to their appreciation and understanding of the unit.
- Overall, they compared the unit favourably to other units at the same University.
- The responses on unit workload and difficulty varied from "medium" to "very hard". The number of hours per week required outside class showed a range of values between 2 and 9 hours, with the mean of 5.3 hours. This figure is not excessive considering that for a 10 credit point unit a student is assumed to study for about 10 hours per week. With 3 hours of class contact per week, the students were expected to spend another 7 hours per week on the unit outside class.
- Approximately half of the respondents thought that the pace of the unit was "about right" and half thought that the pace was "too fast". The

	Question	2003 EIC unit Mean Score out of 9	2004 KDDM unit Mean Score out of 9
Learning/Academic Value	1. You found the class intellectually challenging and stimulating	5.5 61.1%	7.6 84.4%
	2. You have learned something which you considered valuable	6.46 71.8%	7.8 86.7%
	3. Your interest in the unit has increased as a consequence of this class	6.13 68.1%	7.2 80.0%
	4. You have learned and understood the subject material in this class	5.75 64.2%	7.8 86.7%
Examinations / Grading	5. Methods of evaluating student work were fair and appropriate	6.08 67.6%	7.8 86.7%
	6. Assessments/Examinations tested units content as emphasised by staff member	6.08 67.6%	8.2 91.1%
Assignments / Readings	7. Required readings/texts were valuable	6.38 70.9%	7.9 87.7%
	8. Readings, assignments, etc. contributed to appreciation and understanding of the unit	6.21 69.0%	7.9 87.7%
Class Rating	9. Overall, how does the class compare with other classes at this institution	N/A	7.9 87.7%
Workload / Difficulty	10. Unit difficulty, relative to other units, was	6.5 72.2%	6.6 73.3%
	11. Unit workload, relative to other units, was	6.41 71.2%	6.5 72.2%
	12. Unit pace was	5.91 65.7%	6.6 73.3%
	13. Average number of hours per week required outside class	5.2	5.3

Table 4: Student feedback

differences in responses were probably due to different backgrounds of the students. Some students came from more technical computer science courses while several were from relatively non-technical information systems degrees. The latter group experienced some difficulties with understanding of data mining algorithms.

	2003 Emerging Issues in Computing	2004 KDDM
1	<p>EIC: You found the class intellectually challenging and stimulating</p>	<p>KDDM: You found the class intellectually challenging and stimulating</p>
2	<p>EIC: You have learned something which you considered valuable</p>	<p>KDDM: You have learned something which you considered valuable</p>
3	<p>EIC: Your interest in this unit has increased as a consequence of this class</p>	<p>KDDM: Your interest in the subject has increased as a consequence of this class</p>
4	<p>EIC: You have learned and understood the subject material in this class</p>	<p>KDDM: You have learned and understood subject materials in this class</p>
5	<p>EIC: Methods of evaluating student work were fair and appropriate</p>	<p>KDDM: Methods of evaluating student work were fair and appropriate</p>
6	<p>EIC: Assessments tested units content as emphasised by staff member</p>	<p>KDDM: Assessments tested units content as emphasised by staff member</p>
7	<p>EIC: Required readings/texts were valuable</p>	<p>KDDM: Required readings/texts were valuable</p>

Table 5: Distribution of scores for questions 1-7

	2003 Emerging Issues in Computing	2004 KDDM
8	<p>EIC: Readings, assignments, etc. contributed to appreciation and understanding of this unit</p>	<p>KDDM: Readings, assignments, etc. contributed to appreciation and understanding of the unit</p>
9		<p>KDDM: Overall, how does this class compare with other classes in this institution</p>
10	<p>EIC: Unit difficulty relative to other units</p>	<p>KDDM: Unit difficulty, relative to other units</p>
11	<p>EIC: Unit workload, relative to other units</p>	<p>KDDM: Unit workload, relative to other units</p>
12	<p>EIC: Unit pace</p>	<p>KDDM: Unit pace</p>
13	<p>EIC: Average number of hours per week required outside class</p>	<p>KDDM: Average number of hours required outside class</p>

Table 6: Distribution of scores for questions 8-13

7 Reflection

Reflecting back on teaching of the KDDM unit, what aspects of the unit could have been done better? What were the lessons learned?

Overall, the student feedback on the unit was very positive. The only area of concern was the amount of work required which was perceived as excessive by some of the students. In the "additional comments" section of the SEEQ survey, some of the respondents commented that the unit covered too many topics and they had insufficient time to learn all the material. Similar comments were made by several respondents in the 2003 survey.

There are many issues involved in designing a data mining curriculum. The key factors are unit objectives, unit content, and unit assessment. All three must be well coordinated and also must be well matched to the target audience.

7.1 Unit Objectives

Were the unit objectives appropriate? The objectives were chosen to satisfy the core concepts of data mining and ensure a basic understanding of several advanced data mining concepts.

From Table 2 it can be seen that the three continuous assessment items, Assignment 1, Assignment 2, and Project all tested the same six objectives. It would have been better to use finer grained objectives to allow a clear differentiation between the individual assessment tasks. It is important to be able to show how unit objectives are satisfied progressively by different assessment tasks.

Was the range of the unit objectives appropriate? The student feedback and the author's own experience from delivering the two units suggest that similar introductory data mining units should focus their learning objectives towards the core principles and techniques and leave more advanced objectives to later units.

7.2 Unit Content

Was the unit content well chosen? The unit content and unit objectives are closely related. The content must satisfy the unit objectives. Comparing the content topics in the KDDM unit to the ACM SIGKDD guidelines in the draft proposal (SIGKDD, 2006), it can be seen that the KDDM unit corresponded quite closely to the *Foundations (Course I)* unit in the proposed curriculum. However, from students' responses to the evaluation survey, it appears that it would have been better to omit the advanced topics and focus on fewer core topics in more detail.

7.3 Unit Assessment

Were the assessment tasks appropriate? The unit assessment was designed to provide a variety of assessment tasks and to examine as much of the unit content as possible. The assignments and the project provided the students with practical experience in solving

data mining problems and using data mining tools. The final examination tested overall understanding of the unit content.

On the basis of students' feedback, it can be concluded that the assessment tasks were appropriate and helped the students to appreciate and understand the subject matter.

7.4 Technical Level of Data Mining Units

One important issue to consider when designing a data mining unit is the technical level of the unit. A good understanding of data mining algorithms requires a solid background in mathematics, statistics and algorithms and data structures. Such background knowledge is normally provided in computer science degree programs but unfortunately not in many information systems and information technology programs.

As the author's experience with teaching *Emerging Issues in Computing* in 2003 showed, students who don't have adequate background in statistics and computing may experience major problems with understanding mathematical and algorithmic notation and may find technical data mining concepts too difficult to grasp.

To make data mining accessible to students in less technical fields, the data mining units for those fields would have to be focused more on business problem solving and business applications rather than on theoretical aspects of data mining algorithms.

For more technical disciplines such as computer science and engineering, the data mining units should include technical details of data mining concepts and algorithms as well as projects involving industrial or scientific applications of data mining. Computer science students could also be given programming tasks to implement some of the data mining algorithms.

Employment opportunities in data mining range from technical positions requiring knowledge of statistics, computer programming, machine learning and artificial intelligence to non-technical positions in business analysis. There is clearly a need for data mining units and courses targeted to students from different fields.

7.5 Advanced Data Mining Topics

How should the advanced data mining topics be taught? Data mining is nowadays used in many diverse areas of business, industry and research. A single unit cannot cover all aspects of data mining, several units are required. The ACM SIGKDD curriculum draft proposal gives guidelines for two units: *Foundations* and *Advanced Topics* (SIGKDD, 2006).

The *Foundations* unit is focused on the core data mining concepts of data pre-processing, data warehousing and OLAP, association rule mining, classification, clustering, time series and sequence mining, text mining and Web mining, visual data mining, and social impact of data mining.

The proposal for the *Advanced Topics* unit includes advanced material for the core topics discussed in *Foundations*, as well as additional advanced data mining

topics. It contains advanced material on data pre-processing, data warehousing and OLAP, association rule mining, classification, clustering, and time series and sequence mining. In addition, it includes material on data streams mining, spatial, spatiotemporal and multimedia mining, biological mining, text mining, hypertext and Web mining, data mining languages, standards and system architectures, data mining applications, data mining and society, and trends in data mining.

The curriculum proposal does not yet provide any recommendations on how to select appropriate topics to create units for students from different disciplines.

The author's experience with teaching the KDDM unit and the feedback received from the students, indicate that it would be better to limit the material in the *Foundations* unit to the basic core topics and cover them in more detail. The topics such as time series mining, sequence mining, text mining, Web mining, and visual data mining could be included only in the *Advanced Topics* unit. The educators developing advanced data mining units will find the *Advanced Topics* unit an excellent source of content from which to choose a suitable subset of topics.

8 Summary

This paper presented the author's reflection on the experience in developing and delivering a data mining unit to a diverse cohort of students. The main outcomes of this study were:

- The students considered the unit content to be valuable and their interest in data mining was increased by doing this unit.
- The students were satisfied with the unit organisation, delivery, and assessment.
- Unit workload and difficulty were considered too high by some of the respondents to the student evaluation survey.
- Unit objectives were not sufficiently detailed to show how different assessment tasks progressively satisfied different unit objectives.
- Unit content was judged to be too broad, beyond the scope of an introductory unit.
- Unit assessment covered all the unit objectives and was considered to be appropriate.
- Students with inadequate background in statistics and computing experienced difficulties with mastering data mining concepts.

The outcomes of this study suggest the following recommendations for designers of data mining units:

- The curriculum for a postgraduate unit in data mining should be designed in relation to the academic background of the student cohort. Students from technical fields such as computer science and engineering appreciate deep knowledge and understanding of theoretical concepts and algorithms. Students from business oriented fields tend to be interested mainly in

applications of data mining to solving specific business problems. They view data mining from a user's point of view.

- Unit objectives and unit content should be developed in parallel. This would ensure that the unit content satisfies the unit objectives, and also would help to determine appropriate level of detail for the unit objectives.
- Unit content for an introductory data mining unit should focus on the core concepts and cover them in depth rather than cover many concepts in a superficial way.
- Unit assessment should contain a sufficient diversity of tasks to motivate students to learn, and the tasks should correspond to the unit objectives. Each unit objective should be assessed by at least one assessment task.
- Advanced data mining topics should be covered in a separate advanced data mining unit.

9 References

- Chawla, N. V. (2005): Teaching data mining by coalescing theory and applications. *Proc. 35th ASEE/IEEE Frontiers in Education Conference*, Indianapolis, IN, USA, 17-23, IEEE.
- Clementine, SPSS Inc. <http://www.spss.com/clementine/>. Accessed 2 Sept 2007.
- Dunham, M. H. (2003): *Data mining: introductory and advanced topics*. Pearson Education.
- FIMI Repository, <http://fimi.cs.helsinki.fi/data/>. Accessed 2 Sept 2007.
- Han, J., Kamber, M. (2001): *Data mining: concepts and techniques*. Morgan Kaufmann.
- Intensive Working Group of ACM SIGKDD Curriculum Committee: Data mining curriculum: a proposal (Version 1.0) <http://www.sigkdd.org/curriculum.php>. Accessed 2 Sept 2007.
- Musican, D. R. (2005): A data mining course for computer science: primary sources and implementations. *Proc. 37th SIGCSE technical symposium on Computer science education SIGCSE '06*, **38** (1): 538-542, ACM Press.
- Saquer, J. (2007): A data mining course for computer science and non-computer science students. *Journal of Computing Sciences in Colleges* **22** (4):109-114, ACM.
- UCI KDD Archive, Information and Computer Science, University of California, Irvine. <http://kdd.ics.uci.edu/>. Accessed 2 Sept 2007.
- Weka 3: Data Mining Software in Java, University of Waikato, NZ www.cs.waikato.ac.nz/ml/weka/. Accessed 2 Sept 2007.
- Witten, I. and Frank, E. (1999): *Data mining: practical machine learning tools and techniques with Java implementation*. Morgan Kaufmann.