Teaching Data Mining By Coalescing Theory and Applications

Nitesh V. Chawla
Department of Computer Science and Engineering,
University of Notre Dame, IN 46556
nchawla@cse.nd.edu

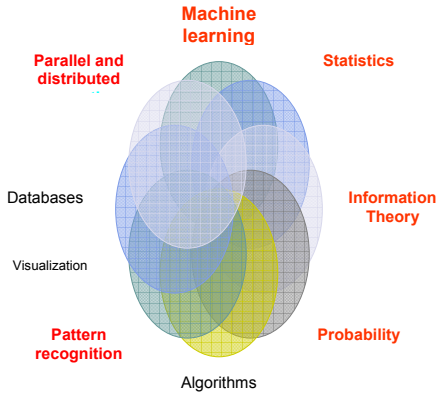
***Abstract* - We report on our experience for the first time departmental offering of the data mining course in Spring 2005. The course was cross-listed such that both the upper level undergraduates and graduate students could attend. However, the majority of the registered students were undergraduates. Data mining, being a confluence of multiple fields, offers an interesting addition to the computer science curriculum. The main objective of the course was to provide grounding on both the theoretical and practical aspects of data mining and machine learning. In addition, the course used concepts learned in various courses throughout the undergraduate degree. The course utilized a machine learning toolkit, Weka, by the University of Waikato, New Zealand. In this paper, we present the various components of the course, structure, innovative assignments and discussions, and the project life cycle.**

Index Terms – Data Mining, Innovative Curriculum, Undergraduate Curriculum

INTRODUCTION

Data mining, being a confluence of multiple fields, offers an interesting addition to the computer science curriculum. It is defined as the "non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" [1]. Data mining draws upon various concepts and courses learned during an undergraduate curriculum, including but not limited to statistics, probability theory, linear algebra, databases, algorithms, and data structures, as shown in Figure 1. This course on data mining was offered for the first time in Spring 2005 at the Department of Computer Science and Engineering at the University of Notre Dame [2]. The main objectives of the course were to:

- Motivate and provide an introduction to the key principles and techniques of data mining;
- Familiarize the students with the complete life cycle of the knowledge discovery process (as typically used in the industry);
- Focus on and encourage analytical and inductive thinking [4, 5];
- Present and discuss some of the key issues in applying data mining to the real-world problems;



- Apply some of the principles and/or techniques learned to different data sets and domains; and

FIGURE 1: CONFLUENCE OF DATA MINING

- Discuss the social aspects of data mining such as the Total Information Awareness [3] project.

The course structure coalesced the theoretical and applied aspects of the field, borrowing motivation from the "real-world" usage of data mining and delving into aspects of the different courses the students have taken during their undergraduate studies. We report on our experience with this course offering, various components of the course, and the students' feedback

Teaching data mining hasn't received a significant amount of attention in the FIE conference series. We are familiar with one paper by Banks et al. [6] at FIE. They provide their experience in offering data mining at two different Universities --- Arizona State and Wayne State University. Their project on teaching data mining in undergraduate engineering was sponsored by the National Science Foundation. We believe and agree with them that data mining is an exciting addition to the curriculum at the senior/graduate level as it offers not only a revision of some of the old concepts, but also offers a stream of computer science education well applied in various domains and applications.

COURSE STRUCTURE

The course followed the key components of the knowledge discovery process. Figure 2 shows the complete knowledge discovery process as per the CRISP data mining standard [7]. Given the applications of the material in the course, we believe it is important to pace the course along a process cycle. This encouraged the students to think along and learn the material as per the different stages in a data mining or knowledge discovery process or project. Establishing this in the beginning of the course also helped the students to plan for their class projects. Figure 3 shows the break-up of the class project.

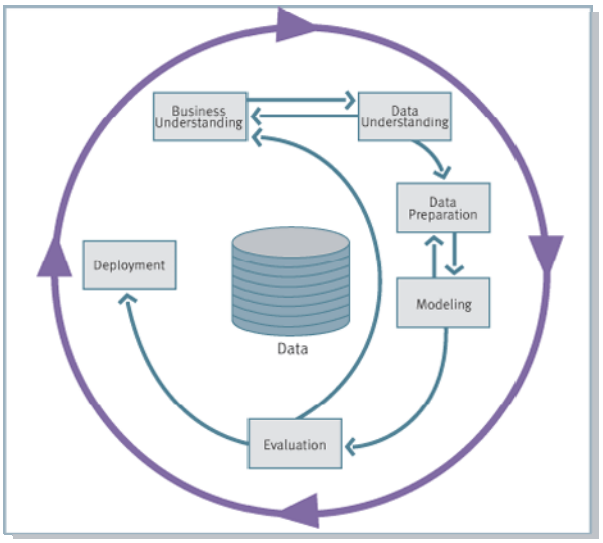


FIGURE 2: CRISP-DM PROCESS CYCLE

The following are the key topics discussed in the course:

1. Introduction to machine learning and data mining
    a. Introduction to data mining (why, how and the inter-disciplinary fertilization).
    b. Brief introduction to machine learning
2. Understanding the data (instances, features, class)
3. Data preparation and preprocessing (missing values, feature selection, noise, etc.)
4. Classification and regression methods
    a. Nearest neighbor
    b. Decision trees
    c. Rule learners
    d. Bayesian learners
    e. Neural networks
    f. Logistic regression
    g. Linear regression and regression trees
5. Unsupervised and semi-supervised learning
6. Evaluation
    a. Concept of validation and testing data
        i. Bootstrapping, 10-fold CV, leave-one-out
    b. Performance metrics (error, ROC, lift, loss functions)
    c. Comparing and benchmarking techniques
7. Ensemble techniques in classification
8. Real-world applications and challenges; emerging and related areas (a discussion)
9. Discussion of role of data mining in society

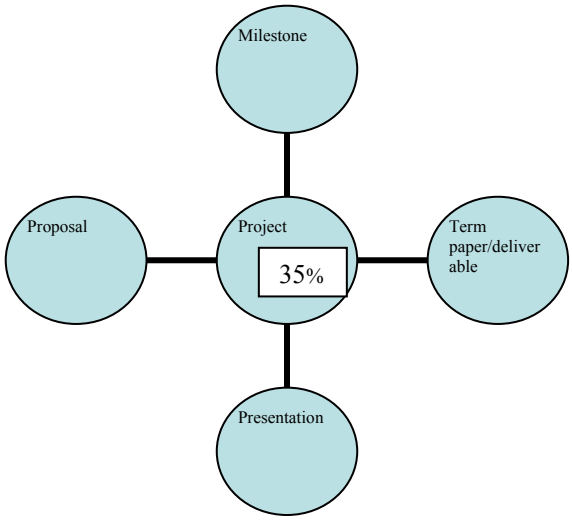


FIGURE 3: PROJECT BREAK-UP AND CORRESPONDING GRADE WEIGHTAGE

As evident by the topics covered, the course indeed invoked various fields and is a compelling addition to the upper-level (junior or seniors) undergraduate electives. The class instruction was done primarily with the use of PowerPoint slides. This allowed integration of multimedia in the course discourse. We used the Data Mining textbook by Witten and Frank [8] as the required textbook for the course. A primary reason for adopting that textbook was the availability of the freely-available and comprehensive tool called Weka [8]. Weka is becoming a popular tool of choice for research and academic discourse. Weka allowed the students to experiment with and analyze different techniques as discussed in the class. The students were favorable about the use of this toolkit. We provide the students' feedback in the subsequent Sections. Moreover, the class projects gave the students freedom to program in their language of choice or add more components to the Weka source code to suit their purpose.

We developed the following course components to teach the various topics:

- **Class lectures**: The class was primarily conducted using PowerPoint slides, which were also provided online as lecture notes**.** Given the inter-disciplinary foci of the field, no single textbook was sufficient. So we prepared all the lecture slides/notes with relevant references and class-handouts.
- **Quizzes and Exams**: There were four quizzes and one midterm, which evaluated the students on all the material taught in-class. The quizzes and midterm evaluated the theoretical understanding of various concepts.
- **In-class Assignments**: Given the confluence of data mining as shown in Figure 1, some of the

mathematical and statistical concepts were not immediately clear to the students. So, we developed various in-class assignments, wherein the students worked together to apply an equation or mathematical formulation to a "toy" problem. The concept was explained at length first, and then the students were given a sheet to work out the toy problem. If a student succeeded, he or she was asked to do it on the blackboard. We then provided the solution to the students, and explained how the concept worked when applied. This really helped the students understand, and develop problem solving and analytical skills. Some of the in-class assignments included information gain, paired-t tests, Bayesian formulations, etc. This exercise prompted the students to work together, apply the concepts immediately, and even encouraged attendance as they were graded on this participation.

- **Homework Assignments**: The homework assignments were divided among theoretical concepts, Weka, and a Matlab assignment. Weka assignments immediately followed lectures on different algorithms and techniques. Thus, practical inferences succeeded the theoretical justifications in-class. The assignments incorporated experimental design, comparison of different methods, statistical tests of significance, and different types of data. The one Matlab assignment provided an opportunity to learn about some of the statistical and visualization methods in Matlab for analyzing data. These included box plots, discretization methods using percentile based binning, histograms, computing sufficient statistics such as mean, variance, etc.
- **Presentation on a real-world application**: We provided various applications of data mining and machine learning to the students' listserv. As part of the in-class assignment and participation, each one of them was asked to prepare any one application for a 10 minute presentation. Each student was then required to comment on each presentation and answer questions such as: *Why data mining was needed? How did it help?* This assignment allowed the students to learn more about an application that deeply interests them, and also become aware of various other applications. In addition, it offered an opportunity for practicing their presentation skills. Moreover, we delivered lectures on various applications, including one invited lecture on Bioinformatics. We also prepared a lecture on credit scoring based on our experience that became among the most popular lectures.
- **Class challenge assignment**: We incorporated this assignment fairly late in the semester. The purpose of the class challenge was to expose the students to the complete process of data mining, and identifying the optimal methods for the dataset at hand. We realized introducing a competitive aspect, and making it require teamwork might just provide the right motivation for the students to give their best. They were simply given a classification and a regression dataset. They were then asked to identify their choice of methods on a validation set, and submit predictions on a testing set. This mimicked a real-world deployment of a data mining system requiring noise cleaning, feature selection, benchmarking various methods on a validation set, and then picking up the best one for the actual testing set predictions. The students were grouped in pairs for this assignment. Based on our experience and students' feedback, we are definitely going to make it a permanent fixture of our course and highly recommend the same for other Universities. This challenge forced the students to think about various aspects of the data, methods, and concepts learned in the class.
- **Data mining and society**: The course also encouraged in-class discussions of the role of data mining in society by providing relevant material from academic literature and popular media. Recent years have included a focus on privacy-preserving data mining due to a public concern about data mining being too invasive. We discussed those topics, including the Total Information Awareness initiative of the Government [3]. We also provided some of the material from Congressional hearings on the data mining technology [12]. We believe it is important to include a discussion on ethics and social impacts in computing [13]. In fact, one of the students is continuing for a degree in Law, and thought this lecture provided an interesting avenue for further exploration for his graduate degree.
- **Class project**: The class project made up 35% of the class grade, as shown in Figure 3. It was subdivided into a proposal, milestone assessment, final deliverable/term paper, peer-review process, and a presentation. The students were required to follow the ACM Knowledge Discovery and Data Mining Conference paper format [9] for writing all the documents during the course of the project. This ensured a standard among all the students, and familiarized them with the conference format of a key data mining conference.

  The students were required to submit a 2-page proposal with what they expect to accomplish in the project. They also included the potential milestones in their proposals. This ensured personal benchmarks, and allowed the students to pace themselves as the semester progressed. Moreover, any road-blocks were identified early on, and alternative paths were considered in advance. For instance, a couple of projects required slight revisions in their goals based on these milestones, and the students were appreciative of this aspect. This can also be typical of a real-world project --- the projects rely on timelines

and milestones. The term paper followed a peer-review process among the students. The goal of this exercise was to acquaint the students with the peer review process. The students conducted a single-blind peer-review process. We provide students' feedback on the peer-review process in the next Section. In addition, the final project presentations were converted into a half-day Workshop. We believe the project structure better prepared the students for various components of academia and industry. We encourage the incorporation of a similar project cycle not only for this course but also for other courses with an applied flavor. Figure 3 shows the distribution of the grade for the class project.

We encouraged the students to consider class projects that aligned well with their career or research interests. We individually defined and discussed each topic of interest, identified the work entailed, and whether the undertaking was doable in a semester. This initiative encouraged a variety of class projects with each student being very enthusiastic about his or her undertaking. In fact, a couple of juniors are continuing with their research projects in data mining over the Summer. A big achievement from the class projects was the acceptance of two project papers (marked by *) at Conferences [14, 15]. To name some class project topics:

1) Spam detection.
2) Intrusion detection.
3) Activity mining in open source software. *
4) Personalized music recommender system.
5) Data mining stock indices.
6) Meta-learning to select classifiers from ensembles. *

Class composition

Table 1 gives the class composition in terms of number of juniors, seniors, and graduate students. In addition the major of each student is also indicated (CS: Computer Science; CE: Computer Engineering; CSE: Computer Science and Engineering). There were a total of 11 (9 for credit and 2 for audit) students in the class.

TABLE 1: CLASS COMPOSITION

Juniors	Seniors	Graduate Students
2 (CS)	3 (CS) 1 (dual CS+MBA) 1 (dual CS + Film)	4 (2 for credit and 2 for audit)

Both the juniors motivated by the course decided to pursue research in data mining (and have published papers from their class projects). One of the graduate students utilized the class to propel his research in intrusion detection, and is continuing the research during Summer. In addition, both the dual major students have noted the class to be very useful for their career objectives in consulting firms. Thus, the students found the course very useful for their varied interests either in academia or industry.

COURSE EVALUATION

In this Section, we present the course evaluation as per the inputs from the students. We will evaluate the following:

- Weka
- Discussion of Applications
- In-class examples/assignments
- Role of applications' discussions in-class
- Project life cycle

Weka is a freely available machine learning toolkit from the University of Waikato [8]. It is a JAVA based toolkit that allows for both the command line interface and the graphical user-interface. Weka was easy to install and learn to use. Most of the students installed Weka on their personal PC's., thus saving them trips to the departmental labs for finishing their assignments

Weka made it easier to conduct homework assignments on different algorithms and techniques, including the class challenge. Moreover, more components can be added to Weka and it can be customized for a particular requirement. We posited the following question to the students for the evaluation of Weka. Note that we only provide a sampling of responses for space reasons.

Do you think Weka is a useful tool for this course? Would you rather program different methodologies?

- "Weka is good; I think its better for us to spend our time gaining a general knowledge of many algorithms than to spend hours becoming intimately familiar with one."
- "I think Weka is a great toolkit, because it allows students to begin exploring without getting in over their heads."
- "Weka is a very useful tool as long as the theories/algorithms are understood."
- "Weka is very useful. I'd rather not do programming assignments -- I'd rather master Weka so that I can apply these skills to applications."
- "Weka is very useful, doing programming forces you to work more time programming and debugging rather than learning and applying new knowledge."
- "Its good to use a toolkit like Weka; it might be interesting to do some programming, but it would be a lot on top of our research assignment."

Summarizing the students' responses, it is clear that the students prefer using a toolkit that allows them to experiment and analyze different techniques for assignments. They wanted to develop an intuition behind different techniques and how they might be used for their jobs and/or research. From an instructor's viewpoint, Weka indeed helps as some of the class discussions can be easily translated into homework assignments not requiring a significant amount of time. However, the students were free to implement the class project in the programming language or environment of their choice.

So that gave them an opportunity to independently develop their ideas and programs.

Our next point of evaluation is the usefulness of application discussions in class. We brought in our real-world experience of applying data mining for various problems. The students found these discussions very useful, and were attentive and keen to understand how the problems were solved. We also provided weekly or biweekly nuggets of some cool applications of machine learning and data mining to the students. We again provide the (undergraduate) responses from the students' survey:

Do you feel the discussions of the real-world perspective and applications are useful?

- "Very much so. I would be a little lost w/o some grounding in real-world applications."
- "Definitely, they motivate learning and many people (myself included) learn best from analyzing applications."
- "Very, Professor has good real-world knowledge."
- "Yes – It provides a good understanding of how the concepts can be used."
- "Yes, the professor's real-world experience allows him to point out the practical aspect of data mining along with common pitfalls."
- "Yes – it is difficult to be enthusiastic about a subject with little real-world application."

Summarizing the students' responses, the applications played an important role in their understanding of the concepts and piqued their interest in the class. We also found that the students were very receptive of the applications, and it allowed them to keep pace with the evolving nature of the machine learning and data mining field. Applications have been a cornerstone in machine learning, as they have identified some of the key challenges and research issues in the area [10].

The next point of course evaluation is the usefulness of in-class assignments. Various data mining and machine learning concepts can be mathematically involved, and confuse the students. We prepared in-class assignments, whenever possible and relevant, for the students to apply the equations to immediately after presentation in the class. The students worked through it together or with hints from the Professor. A complete solution was provided in the end or if one of the students solved it completely, he or she was asked to write it on the black-board for all the students. We provide a feedback from students on that:

Do you find the in-class assignments/examples useful?

- "In-class assignments help very much in understanding the material."
- "Very useful, help nailing down concepts."
- "The examples are very helpful in illustrating the concepts. Applications discussions are very helpful, especially the Professor's real-world experience."
- "Yes, they are great".
- "Yes, I've done most of the assignments by re-working the in-class examples."

From the instructor's perspective, we also found solving an example in class that breaks down the equation into pieces a useful exercise in explaining a concept. It resulted in more time preparing the class material but less time in-class explaining a concept. However, it might not always be possible to do that. So, we noticed that the students were more wary (for some of the techniques) whenever we went "under-the-hood" and provided a bunch of equations. Another key point noticed by the students was that the course required them to synthesize various concepts that were learned over the undergraduate years. Another concept that was slightly difficult for the students was the extensive statistical discussions at times [11].

We also asked the students to evaluate the project breakdown into proposal, milestones, etc. as stated before.

Do you find the project structure as in this Course useful?

- "Yes, I think it will provide good insight into applying DM to an area that we have personal interest in."
- "Yes, even more so than the tests or quizzes."
- "Yeah, it's the same kind of breakdown our projects in real life are going to have."
- "Yes, CS is a major in which understanding the process of a project helps prepare for the future in many ways."
- "Yes, lets us put our education to use with a real-world application."
- "Absolutely, it mirrors professional life so we should be doing more of that."
- "Yes, the deadlines will keep me honest and on track."

Thus, the students were indeed appreciative of this format. Moreover, the upper-class composition of the class may help from a research and real-world exposure to prepare them for their respective career paths. The process of writing proposals, meeting timelines, making presentations, writing papers, and doing peer review will be more than likely present in any career path one chooses. The students were greatly appreciative of the review process as well (4 of them responded that it was very helpful; 3 said it was moderately helpful; and only 1 said that it was of little help).

We believe integrating the various components in the project and spreading them across the entire semester, prepares the students for both academia and industry. We then asked students the question, whether the course was useful in their career plan. Most of them responded with an interest in continuing research and/or job in this area. A few of them also said that this course provided them with something in their CS degree that they could easily apply to the business world.

SUMMARY

We developed a Data Mining course that successfully the course successfully theory and applications. We believe our real-world experience helped in understanding some of the problems when applying data mining to the real-world tasks, and incorporating those in lecture discourse and notes. The course invoked analytical and creative thinking among the students. The course material included both relevant research papers provided as handouts in class and various applications' email nuggets. The students found the use of the applications along with the theoretical framework very relevant. We believe the course can become a very important addition as an undergraduate elective, particularly because of the applied and inter-disciplinary foci. The course synthesized various concepts that students have learned over their undergraduate studies as indicated in Figure 1. As one of the students pointed out, the course made him go back and brush up on all the probability and statistics concepts that he never thought he'd use again. Another student pointed that this course was one of the best he took, as it helped him easily visualize where and how something can be applied. Given the ubiquity of applications of data mining, it helps the students to be exposed to various methodologies before proceeding for graduate school and/or jobs.

We also asked the students about the lectures most preferred. Most of the students enjoyed ensembles because of of being able to combine multiple learning algorithms, and improve the overall performance. Most of the students used ensemble based methods for the challenge assignment. The lecture on credit scoring was one of their favorites. They found neural networks' lectures to be the most difficult due to the various derivations. In addition, the incorporation of real-world applications, examples, project, and the class challenge made the class exciting, although time consuming.

FUTURE WORK

We recommend the usage of Weka as it indeed allows the students an easy evaluation of various techniques. We also recommend an inclusion of the discussion of various applications as it helps the students to attach a tangible element to the various theoretical concepts discussed in the class. We also believe that the break-up of the class project into various components as described in this paper prepares the students for future work. We are going to involve this framework of projects with other course offerings in the subsequent semesters. We believe the students should be ready for the writing and peer-review process that is essential for any research. We hope other Universities include data mining as part of their undergraduate offerings to the upper-classmen. We hope our experience with the course is useful as they plan the course.

We plan to continue usage of this baseline model for our subsequent offerings of data mining, and gather students' responses over different semesters. We will, obviously, improvise the course as it progresses over semesters. One thing that we would like to do is use blackboard based teaching when engaging the students through various mathematical derivations and proofs, such as in neural networks, regression, etc during the course. This will help in pacing the class. We believe a mix of power-point and blackboard teaching is important for a specialized course like this one.

In our subsequent offering, depending on the size of enrollment, we plan to offer another small project as part of the class wherein students will be randomly paired-up. Hopefully, this will imitate the real-world setting of adapting and working with people we have never worked before. And it should also improve the overall chemistry of the class. We would also like to simultaneously conduct the course offering along with our peers at different Universities and jointly monitor the progress of the course. We will share our experience with the community again at that point.

ACKNOWLEDGMENT

We would like to thank the students of the Data Mining Class CSE 498C/598C, Spring 2005 at the University of Notre Dame for their enthusiasm and very useful feedback. We are grateful to Gregory Piatetsky-Shapiro for providing some of the material and his very useful KDnuggets website. We thank the anonymous reviewers for their very useful feedback and comments. We would also like to thank Kevin Bowyer and Larry Hall for their comments on this paper and the course syllabus. We would also like to thank Joaquin Candela for providing the classification dataset that was used in the Challenge.

REFERENCES

[1] Fayyad, U., Piatetsky, G., Smyth. P., "From Data Mining to Knowledge Discovery in Databases," *AI Magazine* 17(3): Fall 1996, 37-54.

[2] Data Mining, Spring 2005, University of Notre Dame, http://www.cse.nd.edu/courses/cse498c/www/

[3] Total/Terrorism Information Awareness Project, http://www.darpa.mil/DARPATech2002/presentations/iao_pdf/slides/PoindexterIAO.pdf

[4] Ford, N. "The growth of understanding in Information Science: Towards a Deveopmental Model", *Journal of the American Society for Information Science,* Vol. 50, No. 12, 1999, pp. 1141-1152.

[5] Trochim, W. M. K., "Deductive and Inductive Thinking", *Deduction & Induction,* http://trochim. human.cornell.edu/kb/dedind.htm, 2002.

[6] Banks, D. L., Dong, G., Liu, H., and Mandvikar, A. "Teaching Undergraduates Data Mining Engineering Programs," *34th ASEE/IEEE Frontiers in Education Conference*, 2004.

[7] CRoss Industry Standard Process for Data Mining, http://www.crisp-dm.org/

[8] Witten, I. , and Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations,* Morgan Kaufmann, 1999.

[9] ACM SIG Proceedings Templates, http://www.acm.org/sigs/pubs/proceed/template.html

[10] Provost, F. and Kohavi, R. "On Applied Research in Machine Learning." Guest editorial in *Machine Learning 30 (2/3) 1998.*

[11] Shield, M. "Statistical Literacy and Mathematical Thinking", 2000 *Presentation at the ICME-9*, Tokyo.

[12] Committee on Government Reform, "Data Mining: Current Applications and Future Possibilities", March 2003, http://www.house.gov/reform.

[13] Bowyer, K. W., *Ethics and Computing: Living Responsibly in the Computer World,* IEEE Press/Wiley Press, 2001.

[14] Mack, D., Chawla, N. V., Madey, G., "Activity Mining in Open Source Software," Accepted in *NAACSOS,* 2005.

[15] Sylvester, J., and Chawla, N. V. "Evolutionary Ensembles: Combining learning agents using genetic algorithms," Accepted in *AAAI Workshop on Multi-Agent Learning,* 2005.