AC 2011-1958: A PRACTICE-ORIENTED APPROACH TO TEACHING UNDERGRADUATE DATA MINING COURSE

Dan Li, Northern Arizona University

Dr. Dan Li received her Ph.D. degree in Computer Science from the University of Nebraska - Lincoln, in 2005. She is currently an Assistant Professor in the Electrical Engineering & Computer Science Department at the Northern Arizona University. Her current research interests include large-scale databases, spatio-temporal data mining, information security, and computer science education.

A Practice-Oriented Approach to Teaching Undergraduate Data Mining Course

Abstract - Data mining is a fast-growing field of study in Computer Science and Information Systems. Many schools have developed data mining course for undergraduate students. The course content has been well defined and streamlined because of the availability of outstanding data mining textbooks. However, the focus on theoretical contents of data mining makes it hard for undergraduate students to digest, and thus, compromises the overall learning outcome. To create an effective and dynamic learning environment, we introduce a practice-oriented approach. This paper describes how we integrate the hands-on component into the course work to enhance the learning of the core data mining topics including data preprocessing, association mining, classification, cluster analysis, text mining, and visualization. The open-source data mining tool, RapidMiner, is introduced to assist students to explore and digest various data mining processes and algorithms. Overall, the hands-on experience provides students a better insight into data mining functions.

1. Introduction

The explosion of very large databases has created extraordinary opportunities for monitoring, analyzing and predicting global economical, geographical, demographic, medical, political, and other processes in the world. Statistical analysis and data mining techniques have emerged for these purposes. Data mining is the process of discovering previously unknown but potentially useful patterns, rules, or associations from huge quantity of data. Typically, data mining functionalities can be classified into two categories: descriptive and predictive. Descriptive mining tasks aim at characterizing the general properties of the data in the databases, while predictive mining tasks perform inherence on the current data in order to make prediction in future [3]. Data mining is not a stand-alone field of study. Instead, it is an interdisciplinary field integrating the components from other fields in Computer Science such as machine learning, pattern recognition, artificial intelligence, visualization, and database systems. Therefore, many Computer Science programs have introduced data mining courses into their curricula.

The course content for data mining has been well defined and streamlined because of the availability of outstanding data mining textbooks [3, 9, 10]. In addition, the ACM SIGKDD Executive Committee has set up the ACM SIGKDD Curriculum Committee to design a sample curriculum that provides guidance and recommendations for educating the next generation of students in data mining. This sample curriculum concentrates on *long-lasting scientific principles and concepts* of the fields [1]. However, the focus on theoretical contents of data mining makes it hard for undergraduate students to digest, and thus, compromises the overall learning outcome. To create an effective and dynamic learning environment, we introduce a practice-oriented approach. This paper describes how we integrate the hands-on component into the course work to enhance the learning of the core data mining topics. The hands-on projects give students an opportunity to carry out experiments that illustrate core concepts in a realistic setting. In addition, the open-source data mining tool, RapidMiner, is introduced to assist students to explore and digest various data mining processes and algorithms.

The rest of this paper is organized as follows. Section 2 describes the background of the data mining course offered in Fall 2010. It lists the core topics covered in this course offering as well as the hands-on experiments to support course objectives. Section 3 describes the practice-oriented methodologies in details focusing on several selected sample projects. Section 4 presents student assessment methodology, the assessment findings, and selected course evaluations. Finally, concluding remarks along with directions for future improvements are presented in Section 5.

2. Course Outcome and Objectives

Data mining is an elective Computer Science course taken by juniors and seniors in Computer Science at Northern Arizona University (NAU). The overall course outcome is outlined in the syllabus as "*Successful completion of this course will provide a student with the necessary skills to design basic data mining algorithms to solve a variety of real-world applications.*" In Fall 2010, we offered this course for the third time after we created this course in Spring 2006. In the first two offerings, we mainly focused on the theoretical contents of data mining with only one team project as a practical experiment. The assessment results were not as good as expected in the first two offerings. Given the applied nature of data mining, we decided to introduce a practice-oriented approach in the third offering to better serve the course outcome. Based on the recommendations from the ACM SIGKDD Curriculum Committee, the objectives of integrating the hands-on component into the theoretical knowledge delivery are five-fold.

a) Learn to use data mining systems by using some data mining software.
b) Implement some data mining functions including association mining, classification, clustering, text mining, and visualization.
c) Implement, refine, and compare of several different data mining methods.
d) Propose, implement, and test new data mining solutions.
e) Use real data to implement and test data mining functions.

To support the above five objectives, we integrate practical components into almost every aspect of the core data mining topics. Table 1 lists all the core topics suggested by ACM SIGKDD Curriculum Committee [1] and shows the topics covered during the course offering in Fall 2010. The table also shows the hands-on experiments associated with the core topics. These experiments are either individual or team experiments, and some of them are required for both. The last column in Table 1 shows the objectives supported by the corresponding hands-on experiments. We can see that each objective is supported by at least three hands-on experiments. This provides students plenty of opportunities to explore, implement, and digest data mining algorithms and functions. Due to time constraint, some core topics suggested by ACM SIGKDD Curriculum Committee are not directly covered in the lectures but are covered through hands-on projects. This stimulates students' self-learning skills and enhances their learning through hands-on experiments.

3. Methodologies

In this section, we describe the practice-oriented methodologies in details focusing on several selected hands-on projects. The course objectives described in Section 2 are addressed and implemented by applying these methodologies.

Table 1: Course Coverage and Hands-on Experiments

Core Topics	Course Coverage	Hands-on Experiment	Objectives Supported
1. Introduction	✔		
2. Data Preprocessing	✔	Individual	a
3. Data Warehousing and OLAP for Data Mining	✗		
4. Association, Correlation, and Frequent Pattern Analysis	✔	Team	d
5. Classification	✔	Individual and Team	b, c, d, e
6. Cluster and Outlier Analysis	✔	Individual and Team	a, b, c
7. Mining Time-Series and Sequence Data	✗	Individual	a, b, c, e
8. Text Mining and Web Mining	✔	Team	d, e
9. Visual Data Mining	✗	Individual	c, d
10. Data Mining: Industry Efforts and Social Impacts	✗		

3.1 Explore the Tool

One of the course objectives is to learn to use data mining systems by using some data mining software. Typical such software may include Microsoft SQLServer 2008 (Analysis Services), IBM Intelligent-Miner, statistical analysis software tools such as R [4], and some open-source data mining tools such as Weka workbench [2] and RapidMiner [7]. RapidMiner (formerly YALE) is one of the most comprehensive and the most flexible data mining and text mining tool. It is the worldwide leading open-source data mining solution due to the combination of its leading-edge technologies and its functional range. RapidMiner and its plug-ins provide more than 400 operators for all aspects of Data Mining. It has a nice and easy-to-use graphical user interface, and it offers many visualizations, pre-processing, machine learning, validation, and automated optimization schemes. Given these helpful features, to enhance students' hands-on experiences, we choose RapidMiner as the tool to explore some data mining processes, algorithms, and functions.

The first warm-up hands-on project is for students to get exposure to the data mining tool, RapidMiner. This tool would be used later in the class to provide students an opportunity to apply a few basic theoretical concepts taught in class.

In this hands-on exercise, students test out two basic operators in RapidMiner, *Retrieve* and *Replace Missing Values*. The *Retrieve* operator can be used to access the repositories introduced in RapidMiner. It should replace all file access, since it provides full meta data processing, which eases the usage of RapidMiner a lot. In contrasting to accessing a raw file, it will provide the complete meta data, so that all meta data transformations are

possible. The *Replace Missing Values* operator replaces missing values in the data set. If a value is missing, it is replaced by one of the functions *minimum*, *maximum*, *average*, and *none*, which is applied to the non-missing attribute values of the data set. We choose this operator because missing data manipulation is one of the most important steps in data preprocessing.

Students are required to submit two screenshots to show the meta data of a given data set before and after applying the *Replace Missing Values* operator, respectively. Figures 1 and 2 show the results generated by RapidMiner.

Name	Type	Statistics	Missings
class	nominal	mode = good (26), least = bad (14)	0
duration	integer	avg = 2.103 +/- 0.754	1
wage-inc-1st	real	avg = 3.621 +/- 1.331	1
wage-inc-2nd	real	avg = 3.913 +/- 1.281	10
wage-inc-3rd	real	avg = 3.767 +/- 1.415	28
col-adj	nominal	mode = none (14), least = tcf (4)	16
working-hours	integer	avg = 37.811 +/- 2.717	3
pension	nominal	mode = none (8), least = ret_allw (3)	22
standby-pay	integer	avg = 6.143 +/- 4.845	33
shift-differential	integer	avg = 4.583 +/- 4.754	16
education-allowance	nominal	mode = no (11), least = yes (7)	22
statutory-holidays	integer	avg = 11.105 +/- 1.371	2
vacation	nominal	mode = below-average (14), least = average (11)	3
longterm-disability-assistance	nominal	mode = yes (11), least = no (5)	24
contrib-to-dental-plan	nominal	mode = half (11), least = none (6)	15
bereavement-assistance	nominal	mode = yes (18), least = no (2)	20
contrib-to-health-plan	nominal	mode = full (12), least = none (6)	16

Figure 1: Meta data before applying the *Replace Missing Values* operator.

Name	Type	Statistics	Missings
class	nominal	mode = good (26), least = bad (14)	0
duration	integer	avg = 2.103 +/- 0.744	0
wage-inc-1st	real	avg = 3.580 +/- 1.339	0
wage-inc-2nd	real	avg = 4.685 +/- 1.747	0
wage-inc-3rd	real	avg = 3.767 +/- 0.752	0
col-adj	nominal	mode = none (30), least = tcf (4)	0
working-hours	integer	avg = 37.811 +/- 2.610	0
pension	nominal	mode = none (30), least = ret_allw (3)	0
standby-pay	integer	avg = 6.143 +/- 1.900	0
shift-differential	integer	avg = 4.583 +/- 3.651	0
education-allowance	nominal	mode = no (33), least = yes (7)	0
statutory-holidays	integer	avg = 11.105 +/- 1.336	0
vacation	nominal	mode = below-average (17), least = average (11)	0
longterm-disability-assistance	nominal	mode = yes (35), least = no (5)	0
contrib-to-dental-plan	nominal	mode = half (26), least = none (6)	0
bereavement-assistance	nominal	mode = yes (38), least = no (2)	0
contrib-to-health-plan	nominal	mode = full (28), least = none (6)	0

Figure 2: Meta data after applying the *Replace Missing Values* operator.

3.2 Use Real Data

Computer Science is one of the fastest growing fields of study. To lead to a successful career in the CS profession, a student should not only keep up with the trends of knowledge development, but also, most importantly, be able to apply the knowledge to analyze and solve real-world problems. Based on the past experiences, students responded best when they thought of class work as an opportunity to work on authentic problems rather than as an obligation to fulfill a class assignment. Therefore, one of the course objectives is to use real data to implement and test data mining functions for real-world applications.

To meet this end, we choose one task in the 2001 KDD Cup data mining competition, i.e., prediction of gene localization [6]. Using the real data for this project is a great challenge. From data mining point of view, the important challenge is to find a way to efficiently use the *Interaction_relation* data files, which is not obvious. Another issue is that there is a high proportion of missing variables in the *Genes_relation* data. Students are required to implement one of the classification algorithms covered in class to predict gene localization.

Most students use K-Nearest-Neighbor (KNN) algorithm [3] to tackle this task because this is the approach used by the winner of the 2001 KDD Cup. Figure 3 shows one key screenshot of the interface presenting the training and testing procedures, and Figure 4 shows the accuracy of the predicted results with regard to each available class in the training data set. The overall accuracy generated by one submission is 52.69% as shown in Figure 4. Among all the submissions, the best one has the accuracy of 62.83%, which is 10% less than the winner of the 2001 KDD Cup (who had 72.17% accuracy). With very limited class time, this result is totally reasonable and acceptable.

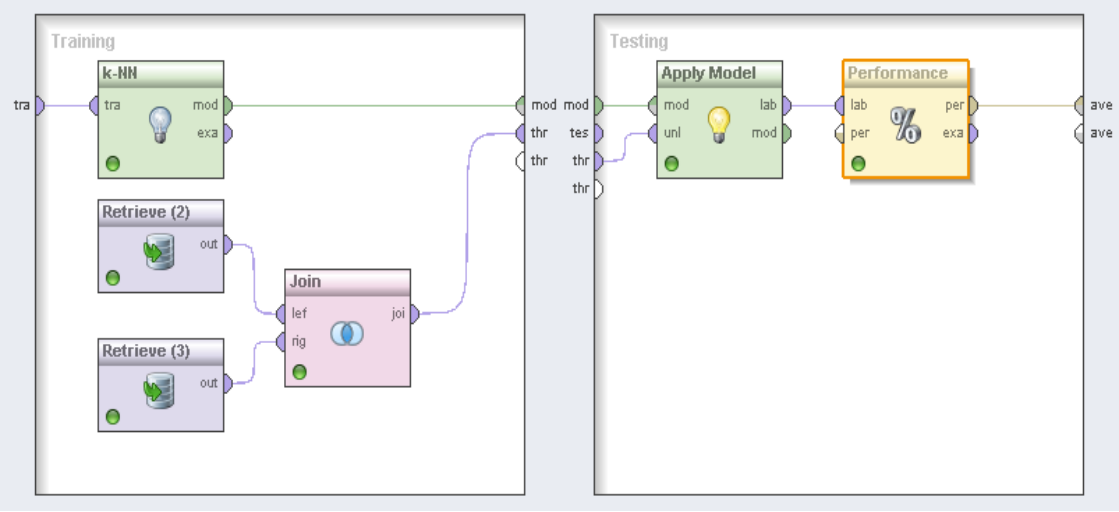

Figure 3: Training and testing procedures using RapidMiner.

accuracy: 52.69% +/- 0.10% (mikro: 52.69%)

	true mitocho	true nucleus	true plasmai	true golgi	true cytoskel	true cytoplas	true vacuole	true er
pred. mitoch	438	40	0	0	0	0	40	0
pred. nucleu	230	5900	0	0	200	184	90	70
pred. plasm:	0	0	0	0	0	0	0	0
pred. golgi	30	20	110	270	0	140	40	30
pred. cytoske	20	90	0	0	720	120	0	30
pred. cytopla	972	2250	550	420	390	2846	330	870
pred. vacuol	0	40	0	0	0	0	0	0
pred. er	0	0	0	0	0	0	0	0

Figure 4: Evaluation of predicted results.

3.3 Combine to Compare

Even though both Computer Science programs and Information Systems programs offer data mining courses, they usually emphasize different topics with different focus. Computer Science programs focus on the thorough understanding of the mathematical aspects of data mining algorithms and the efficient implementation of algorithms, while Information Systems programs focus on the data analysis and business intelligence aspects of data mining [5]. Therefore, one of the course objectives of teaching data mining for Computer Science students is to implement, refine, and compare several different data mining algorithms including classification, clustering, text mining, and visualization, etc.

To support this objective, we develop a hands-on course project to implement and compare at least two clustering algorithms and to visualize the corresponding clustering results. The data set we use is the synthetic control time series data available in UCI machine learning database [8]. It resembles real-world information in an anonymized format. It contains six different classes (Normal, Cyclic, Increasing trend, Decreasing trend, Upward shift, Downward shift). With these trends occurring on the input data set, the clustering algorithm will cluster the data into their corresponding class buckets.

Figure 5 shows the visualization of the clustering result from one student's implementation. Figure 6 shows the visualization of the clustering result generated by RapidMiner. In these two figures, the horizontal axis shows the time span, and the vertical axis shows the synthetic control data over the time. By comparison, the graphs in Figure 6 show a little better clustering result, especially for the downward and the upward trends in the first row. The student provides the detailed analysis in his project report explaining why his implementation does not beat the result generated by RapidMiner. One reason is that the clustering algorithm he implemented, the *k-means clustering*, yields only a local optimum, and thus the final partition might be quite dependent on the initial seeds.

In sum, this project not only requires students to have the deep understanding of theoretical data mining algorithms, it also helps students analyze the performance of their implementations through visualization.

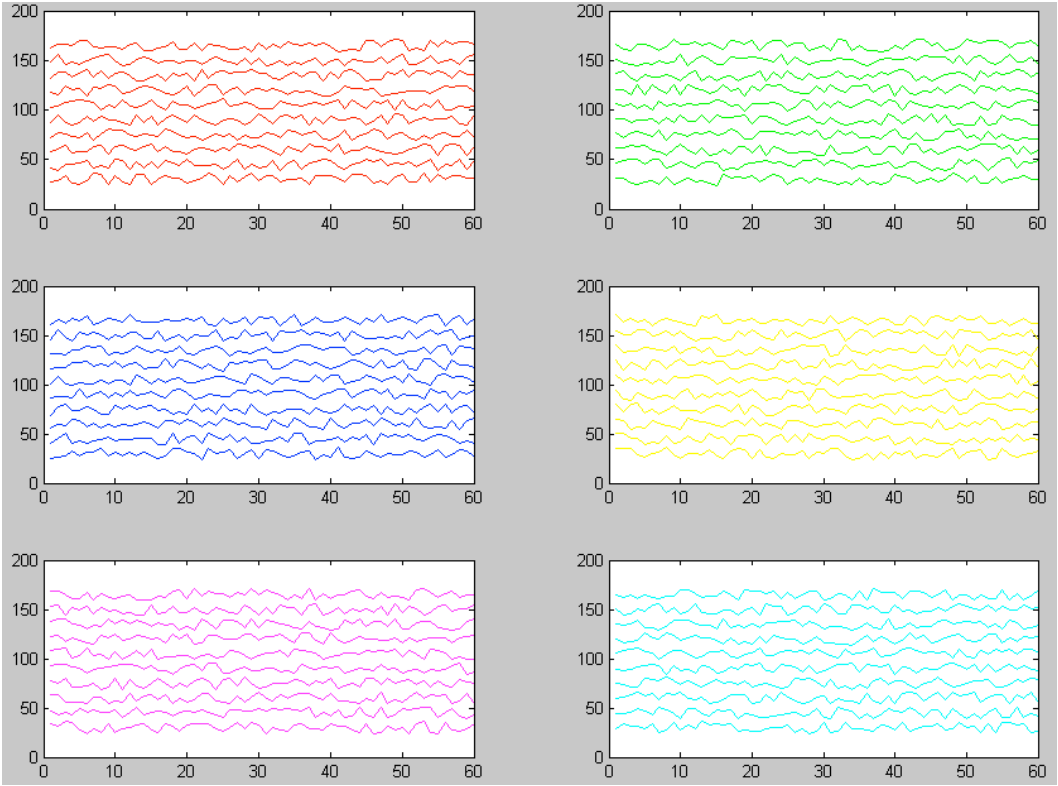

Figure 5: Visualization of clustering result generated by student program.

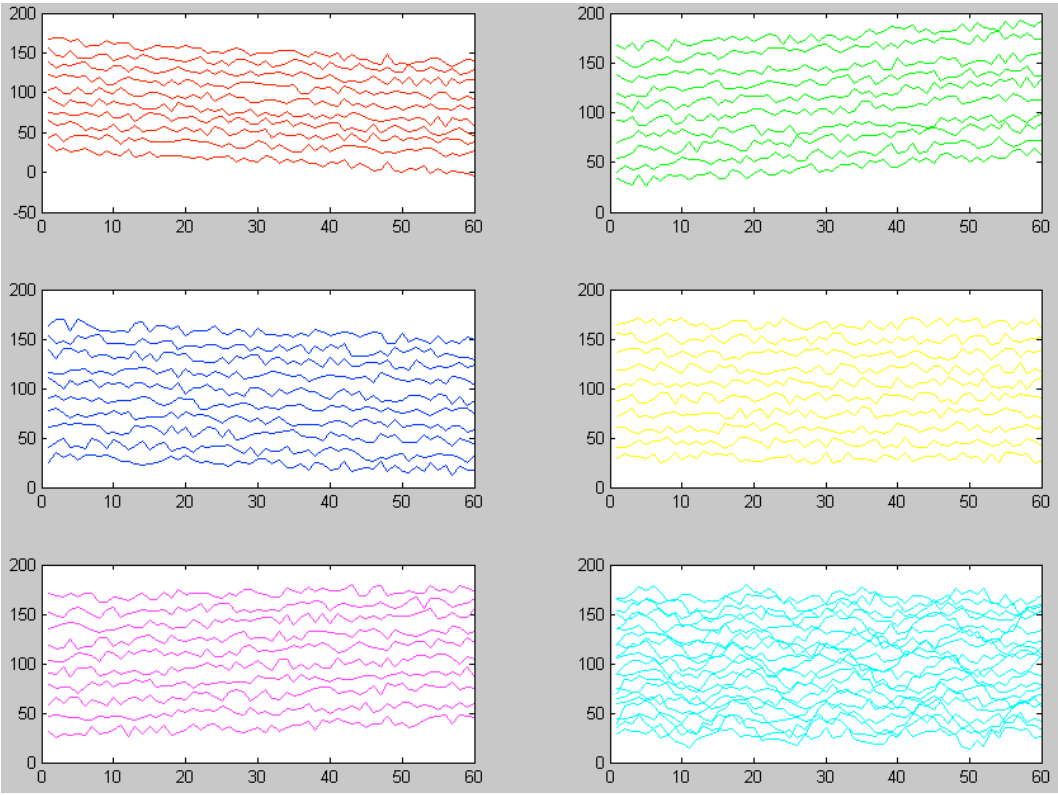

Figure 6: Visualization of clustering result generated by RapidMiner.

3.4 Free of Choice

To generate the greatest productivity and self-satisfaction, we allow students in the data mining class to propose, implement, and test data mining solutions for any projects they have strong passion for. These projects are semester-long team projects. Teams have two weeks to submit their initial project proposals. They can pick the projects suggested by the instructor or propose the new ones. Within one week, they get feedback from the instructor to further refine their proposals. Once the instructor approves their proposals, they can start working on the projects immediately. To make sure they put sufficient effort into the projects throughout the entire semester, they are required to submit project progress reports in the middle of the semester. Finally, by the end of the semester, teams present their works and submit their final reports in technical writing format.

Among six project teams, two of them picked the projects from the list suggested by the instructor, and the other four groups proposed their own projects in the areas where data mining could be well applied. As an example, one group chose to data mine the popular online real-time strategy (RTS) game *StarCraft II* in an attempt to uncover patterns among the decisions made by top-level players of the game. As another example, one project team developed a music search engine by scanning popular RSS feeds from websites such as Reddit.com. In addition to compiling many music results from multiple feeds, the results were ranked based on popularity contributed by multiple metrics.

Another interesting project proposed by a project team was to develop an application toolkit to demonstrate how artificial neural networks can solve classification problems. Instead of applying or developing novel data mining algorithms, this project aimed at helping people understand the underling principles of artificial neural networks through graphic user interface (GUI). Figure 7 shows a snapshot of the application presenting the training, validation, and test errors in network training process.

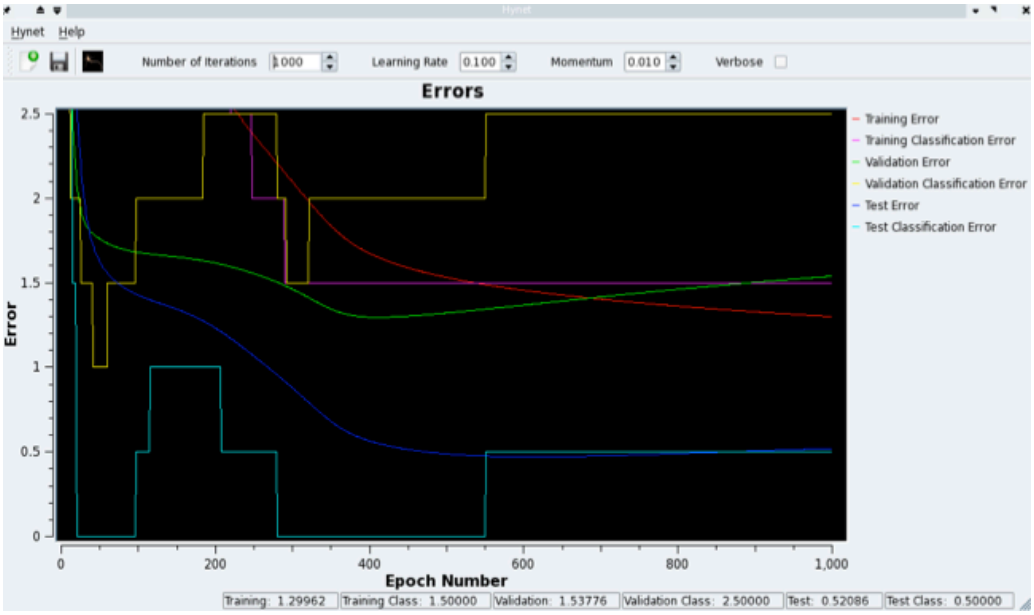

Figure 7: Training, validation, and test errors in network training process.

4. Course Assessment

The overall goal of introducing the practice-oriented approach to teaching data mining course is to create an effective and dynamic learning environment to enhance the learning of the core data mining topics. Therefore, we use the scores the students received on core topics to evaluate the effectiveness of this teaching and learning methodology. The core topics are assessed through theoretical questions including true/false questions, short answers, and algorithm simulations. Table 2 shows the comparison of student scores on core topics in Fall 2008 and Fall 2010 course offerings, respectively. From the table, the improvement on student scores is tremendous. The D&F rate has dropped from 50% to 21%. Even though with the limited number of samples we cannot simply conclude that the practice-oriented approach is the mere contributor to this improvement, without any doubt, the hands-on experiments do help students digest the core theoretical data mining concepts. In future, we plan to develop a detailed assessment rubric to evaluate the effectiveness of course delivery systematically.

Table 2: Comparison of Student Scores

Letter Grade	2008		2010	
	# of student	%	# of student	%
A	3	30	2	14.3
B	2	20	7	50
C	0	0	2	14.3
D	3	30	1	7.1
F	2	20	2	14.3

In addition, the student evaluations also reflect the effectiveness of the practice-oriented teaching strategy. Here are some selected student comments. "*The programming assignments were pretty good.*" "*I enjoyed the projects.*" "*Great feedback, very reasonable expectations, and great practical assignments to implement data mining algorithms and techniques, and to tackle data mining problems of particular interest to the students themselves.*"

5. Conclusions

Data mining is the process of discovering previously unknown but potentially useful patterns, rules, or associations from huge quantity of data. It is an interdisciplinary field integrating the components from other fields in Computer Science such as machine learning, pattern recognition, artificial intelligence, visualization, and database systems. Many Computer Science programs have introduced data mining courses into their curricula. The course content for data mining has been well defined and streamlined because of the availability of outstanding data mining textbooks and the guidance and recommendations provided by the ACM SIGKDD Curriculum Committee. However, the focus on theoretical contents of data mining makes it hard for undergraduate students to digest, and thus, compromises the overall learning outcome.

To create an effective and dynamic learning environment, we introduce a practice-oriented approach to integrate hands-on components into the course work. By doing the hands-on projects, students have their very first opportunity to deal with real-world data

and to carry out experiments that illustrate topics in a realistic setting. The open-source data mining tool, RapidMiner, is introduced to assist students to explore and digest various data mining processes and algorithms. To meet course objectives, we integrate practical experiments into almost every aspect of the core data mining concepts. Each course objective is supported by at least three hands-on experiments. This provides students plenty of opportunities to explore, implement, and digest data mining algorithms and functions. To generate the greatest productivity and self-satisfaction, we allow students in the data mining class to propose, implement, and test data mining solutions for any projects they have strong passion for.

The comparison of student scores received on core topics from the last two course offerings shows the dramatic improvement on students' performance. The D&F rate has dropped from 50% to 21%. Even though we cannot simply conclude that the practice-oriented approach is the mere contributor to this improvement, without any doubt, the hands-on experiments do help students get a better insight into data mining core concepts.

In future, besides RapidMiner, we plan to explore more statistical analysis and data mining tools for algorithm extension and application exploration. Additionally, even though students like the idea of proposing data mining problems of particular interest to them, they need more guidance on selecting appropriate topics and writing the proposal in a professional way. We also plan to develop a detailed assessment rubric to evaluate the effectiveness of course delivery systematically.

References

[1] S. Chakrabarti, M. Ester, U. Fayyad, J. Gehrke, J. Han, S. Morishita, G. Piatetsky-Shapiro, W. Wang, "Data Minig Curriculum: A Proposal (Version 1.0)". Intensive Working Group of ACM SIGKDD Curriculum Committee, April 30, 2006. [Online] http://www.sigkdd.org/curriculum/CURMay06.pdf. [Accessed: 10-Jan-11].

[2] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," SIGKDD Explorations, vol. 11, no. 1, 2009.

[3] J. Han and M. Kamber, "Data Mining Concepts and Techniques". Morgan Kaufmann, 2006.

[4] R. Ihaka and R. Gentleman, "R: A language for data analysis and graphics," Journal of Computational and Graphical Statistics, **5**, 299-314, 1996.

[5] M. Jafar, "A Tools-Based Approach to Teaching Data Mining Methods," Journal of Information Technology Education: Innovations in Practice, Volume 9, IIP-1 – IIP 24, 2010.

[6] KDD Cup 2001. [Online] http://pages.cs.wisc.edu/~dpage/kddcup2001/. [Accessed: 14-Jan-11].

[7] RapidMiner. [Online] http://rapid-i.com/. [Accessed: 13-Jan-11].

[8] Synthetic Control Time Series Dataset. [Online] http://archive.ics.uci.edu/ml/databases/synthetic_control/synthetic_control.data.html. [Accessed: 13-Jan-11].

[9] P. Tan, M. Steinbach, V. Kumar, "Introduction to Data Mining". Addison Wesley, 2006.

[10] I. H. Witten and E. Frand, "Data Mining, Practical Machine Learning Tools and Techniques". Morgan Kaufmann, 2005.