Teaching Data Mining in the Era of Big Data

Dr. Brian R. King, Bucknell University

Brian R. King is an Assistant Professor in computer science at Bucknell University, where he teaches introductory courses in programming, as well as advanced courses in software engineering and data mining. He graduated in 2008 with his PhD in Computer Science from University at Albany, SUNY. Prior to completing his PhD, he worked 11 years as a Senior Software Engineer developing data acquisition systems for a wide range of real-time environmental quality monitors. His research interests are in bioinformatics and data mining, specifically focusing on the development of methods that can aid in the annotation, management and understanding of large-scale biological sequence data.

Prof. Ashwin Satyanarayana, New York City College of Technology

Dr. Satyanarayana serves as an Assistant Professor in the Department of Computer Systems Technology at New York City College of Technology (CUNY). He received both his MS and PhD degrees in Computer Science from SUNY – Albany in 2006 with Specialization in Data Mining. Prior to joining CUNY, he was a Research Scientist at Microsoft, involved in Data Mining Research in Bing for 5 years. His main areas of interest are Data Mining, Machine Learning, Data Preparation, Information Theory, Applied Probability with applications in Real World Learning Problems. Address: Department of Computer Systems Technology, N-913, 300 Jay Street, Brooklyn, NY-11201.

Teaching Data Mining in the Era of Big Data

Abstract

The amount of data being generated and stored is growing exponentially, owed in part to the continuing advances in computer technology. These data present tremendous opportunities in data mining, a burgeoning field in computer science that focuses on the development of methods that can extract knowledge from data. Recent studies have noted the rise of data mining as a career path with increasing opportunities for graduates. These opportunities are not only available in the private sector; the U.S. government has recently invested $200 million in "big data" research. These suggest the importance for us to teach the tools and techniques that are used in this field.

Data mining introduces new challenges for faculty in universities who teach courses in this area. Some of these challenges include: providing access to large real world data for students, selection of tools and languages used to learn data mining tasks, and reducing the vast pool of topics in data mining to those that are critical for success in a one-semester undergraduate course.

In this paper, we address the above issues by providing faculty with important criteria that we believe have high potential for use in an undergraduate course in data mining. We first discuss a set of core topics that such a course should include. A set of practical, widely-accepted tools and languages used for data mining are summarized. We provide a list of sources for real-world datasets that can be useful for possible course assignments and projects. We conclude with a discussion of challenges that faculty should be aware of, including those that were encountered in our course, with suggestions to improve course outcomes. Our paper is based on our collective research and industry experience in data mining, and on the development of an undergraduate course in data mining that was taught for the first time in 2011.

1. Introduction

Enormous amounts of data are generated every minute. Some sources of data, such as those found on the Internet are obvious. Social networking sites, search and retrieval engines, media sharing sites, stock trading sites, and news sources continually store enormous amounts of new data throughout the day. For example, a recent study estimated that every minute, Google receives over 2 million queries, e-mail users send over 200 million messages, YouTube users upload 48 hours of video, Facebook users share over 680,000 pieces of content, and Twitter users generate 100,000 tweets[1]. Some sources of data are not as obvious. Consider the vast quantity data collected from sensors in meteorological and climate systems, or patient monitoring systems in hospitals. Data acquisition and control systems, such as those found in cars, airplanes, cell towers, and power plants, all collect unending streams of data. The healthcare industry is inundated with data from patient records alone. Insurance companies collect data for every claim submitted, fervently working to catch increasing quantities of fraudulent claims. Regardless of the source of the data, contained within them are nuggets of

knowledge that can potentially improve our understanding of the world around us. The challenge before us lies in the development of systems and methods that can extract these nuggets.

We are in a new era in modern information technology - the "Big Data" era. In March, 2012, the U.S. Government announced a "Big Data Research and Development Initaitve" -- a $200 million dollar commitment to improve our ability to "extract knowledge and insights from large and complex collections of digital data." Government agencies such as NSF, NIH, and DoD are investing hundreds of millions of dollars toward the development of systems that can help them extract knowledge from their data.

The career potential for our graduates continue to blossom in this field. A recent study released by Gartner projects that in 2013, "big data is forecast to drive $34 billion of IT spending," with a total of $232 billion to be spent through 2016[2]. In another report, they estimate that "by 2015, 4.4 million IT jobs globally will be created to support big data" with the US generating 1.9 million of those jobs[3]. However, as numerous sources have suggested in recent years, despite the rapid increase in opportunities for careers in big data, there is a dearth of talent. A recent report from the McKinsey Global Institute states that "there will be a shortage of talent necessary for organizations to take advantage of big data. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions."[4]

The topic of "data" is no longer a subject for only the computer science major. As educators, we need to encourage and engage our students in all STEM fields in a way that raises their awareness of the challenges and opportunities in being able to work with large datasets. We need to work hard to provide our students with the knowledge required to work competently in the emerging field of Big Data.

This paper presents an overview of the field of Big Data, focusing primarily on the topic of data mining as the key component to teach our students interested being a part of this field. We present the background and terminology that is pervasive in literature and media; an approach to teaching data mining at an introductory level for undergraduates is presented, including a list of fundamental topics to teach and a set of tools to present that could be used in any data mining course, at any level of student background and expertise; we conclude by presenting a discussion of issues and challenges that we believe educators should be aware of before incorporating the topic of data mining into their course, or introducing a complete course on this important topic into their curriculum.

2. Background and Terminology

There have been numerous names and phrases that have been used to represent different facets of Big Data, many of which are predecessors to this popular term. We will clarify these terms, which are often used interchangeably in the discipline.

Big Data is a catch phrase that describes aspects of the data itself. IBM, a major player in this field, has four descriptors that are used to determine if data are classified as Big Data[5]:

Volume	The sheer size of the data set is enormous, making traditional data processing methods impossible.
Variety	The data can be represented in a wide range of types, structured or unstructured, including text, sensor, streaming audio or video, or user-click streams, to name a few.
Velocity	Many data need to be processed and analyzed in near real-time. For example, consider analyzing stock trades for a sudden move, or catching server attacks.
Veracity	The resulting information for the analyses need to be accurate and trustworthy. This is a huge problem considering the wide range of sources that data comes from. Multiply this to the enormous number of methods that continue to be introduced for data mining purposes, and you have a real challenge in earning the trust of any resulting analysis.

There are many other related fields that have provided an important, long-standing foundation for this area. The broad topic of *database management systems* (DBMS) focuses on the collection, storage, management and retrieval of data. It has become increasingly common for businesses to have multiple databases, from multiple sources, often with their own formats. A *data warehouse* is a type of database that focuses on the aggregation and integration of data from multiple sources, usually for analysis and reporting purposes.

Many fields in Big Data focus on the extraction of information. For example, *Business Intelligence* (BI) systems focus on providing historical, current, and predictive views for the business making use of it. Often, BI and related systems manage their data in the form of *data cubes*, which are multi-dimensional views of data managed and modeled in a way for rapid query and analysis. *Online analytical processing*, or OLAP, is an important part of BI systems that focuses on creating views and queries from data cubes for the purposes of analyzing business data. *Data visualization* is an important field in Big Data that focuses on the development of methods that can help analysts visualize interesting patterns or relationships in data.

Perhaps that most notable term that has been used in this field (and the one that the majority of this paper will focus on) is *data mining*. Used synonymously with the phrase, *knowledge discovery from data*, or KDD, data mining focuses on the development of methods that can automatically extract interesting patterns from data[6], with the hope that these patterns will help the analyst discover new nuggets of information and knowledge from their data. Data mining focuses on methods that can work with large datasets, with the understanding that, for most small datasets, one only needs simple statistical analysis software or a spreadsheet to analyze the data.

3. Data Mining for Undergraduates

Big Data is not tied to any single discipline. Almost all of the STEM fields involve studies where large datasets are collected and analyzed for research purposes. Therefore, it should be no surprise to understand how interest in this field comes from more than computer science students. In our case, the course was first offered as an elective in the computer science (CS) department, and was announced as an introduction to the field of data mining. For its first offering, the class was fully enrolled with 22 students (despite the 20 student cap), with 2 students waitlisted. We had 10% of the enrollment coming from non-CS majors. The following

offering (Spring 2013) has observed a substantially larger interest in the course. The course cap was increased from 20 to 25, with 15 students placed on the waiting list. The interest from non-CS majors grew as well, with 15% of the interest in the course coming from non-CS majors. The instructor for the course had many other non-majors that came to talk about the course, as they were concerned about the computer science prerequisites. Being that CS majors need electives to satisfy their requirements in our program, they were given higher priority with enrollment.

It should come as no surprise that there is a growing interest in the role of data among our student bodies. Students are increasingly aware of the impact of data in their world, regardless of discipline. However, the course designer must consider the trade-offs when allowing non-CS majors with little programming background into the course. You will have increased enrollments, but you must broaden the types of assignments and projects that can be allowed; you also must consider the limitations that may be in place with the algorithms covered for the methods.

Like the data itself, the field of data mining is interdisciplinary. The core of the field relies heavily on the study of data structures and algorithms from computer science, and probability and statistics from mathematics. The field focuses on the application of numerous methods that came out of research in machine learning -- a field whose focus is to develop new algorithms that can "learn," i.e., improve some performance metric, from data. While machine learning tends to focus on theory and algorithm development, data mining focuses on the application of these methods toward large-scale data. Data mining would not have made the great strides it has today without the accomplishments of machine learning.

Depending on how the course is taught, it can be fine tuned for a wide range of audiences by focusing on specific types of data for study purposes. Fortunately, in the era of Big Data, there are very few fields that lack data for analysis purposes. For example, an introductory course can focus on business management by giving exercises incorporating sales and transaction data; it can focus on the biologist or health science student by discussing the wide range of interesting studies analyzing biological or biomedical data (i.e. bioinformatics); it can focus on the engineering student by incorporating exercises that focus on the analysis of sensor data from various instrumentation and data acquisition systems, environmental models of climate, accelerometers mounted on robots or humans for motion analysis, or sensor networks.

3.1 Outcomes

We designed our course with two specific goals in mind. We believe that the there are many careers outside of computer science, particularly STEM-based careers, that can benefit from hiring employees with data mining skills. Therefore, our first and primary goal is to prepare students to successfully work in an entry level position where data mining would be an integral part of their career. A secondary objective is to provide adequate preparation for students that have aspirations for graduate study in data mining. We selected topics in data mining that represent the core material that we believe every data miner should have knowledge of. The topics selected can be covered in a single semester, undergraduate level course.

3.2 Prerequisites

As an elective in the computer science program, we set the most important prerequisite to be our course in data structures. We also expect that students will have a fundamental knowledge in statistics and probability. We had no other prerequisites in place. If you are designing a course solely for computer science majors, then we suggest making a course in algorithm design an additional prerequisite; this will allow you to cover methods with more depth, including critical analyses of computational resource requirements.

3.3 Topics

Here, we present a set of topics that form the basis of a introductory data mining course that will prepare the student to be competent in the world of Big Data. Our list is largely based on a proposed data mining curriculum assembled by a task force that was put together by the ACM SIGKDD curriculum committee and fine tuned based on our own experience[7].

1. Introduction.
Basic concepts, definitions and terminology of data mining should be covered here. Motivate the field with real-world applications of where data mining has been used. We encourage you to discuss current uses of data mining by pulling examples from various news repositories. Discuss some of the implications of data mining, such as privacy and ownership. Using the examples from the news and current events, discuss the different kinds of data repositories on which data mining can be performed. Encourage them to think about the different kinds of patterns and knowledge that can be mined. Introduce the idea of what it might mean for a pattern to be "interesting".

2. Data Preprocessing.
Real world data are generally (a) incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data (b) noisy: containing errors or outliers and (c) inconsistent: containing discrepancies in codes or names. Certain basic preprocessing techniques are discussed here, including:

- Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
- Data integration: using multiple databases, data cubes, or files.
- Data transformation: normalization and aggregation.
- Data reduction: reducing the volume but producing the same or similar analytical results.
- Data discretization: part of data reduction, replacing numerical attributes with nominal ones

3. Association, correlation, and frequent pattern analysis.
Association rule mining is a data mining task that aims to discover attributes in high-dimensional data that have a high probability of co-occurrence in the data. These relationships are discovered through algorithms for frequent pattern analysis. Association rules are often used to analyze sales transactions, and this is the predominant means of teaching this topic.

4. Classification.
Classification is a data mining function that assigns items in a collection to target categories or classes. Falling under the field of supervised learning, or learning from labeled data, the goal of classification is to develop a computational model from existing data that can accurately predict the target class for each new datum that is yet to be observed. A classification task begins with a data set in which the class assignments are known, called labeled data. You will need to cover metrics for measuring classifier performance. For classification methods, we suggest that you cover decision tree algorithms, and Naïve Bayes classifiers, as both are relatively easy to understand. We recommend you discuss automated text classification when discussing Naïve Bayes, as students can understand why this would be important to have, it is easy to understand, and still performs well by today's standards. However, a wide range of examples for classification exist, and you should choose a dataset or two for in-class examples.

5. Cluster and Outlier Analysis.
Here we describe clustering, the unsupervised mining function for discovering natural groupings in the data. It falls under unsupervised learning because the data is unlabeled. Clustering analysis finds clusters of data objects that are similar in some sense to one another. The members of a cluster are more like each other than they are like members of other clusters. The goal of clustering analysis is to find high-quality clusters such that the inter-cluster similarity is low and the intra-cluster similarity is high. Clustering can also be used for anomaly detection. Once the data has been segmented into clusters, you might find that some cases do not fit well into any clusters. These cases are anomalies or outliers.

6. Data Mining: Industry efforts and social impacts[12]
Over the last three decades, increasingly large amounts of critical business data have been stored electronically and this volume will continue to increase in the future. Despite the growing volume of data collected by businesses, few have been able to fully capitalize on its value. This is due to the difficult task of fully analyzing these data and discerning the underlying patterns that can emerge.

For example, most large retail companies collect large quantities of information from every buyer that comes through the store. Retailers can use their customer transaction data to answer a wide range of questions, and ultimately improve their business. Data mining could be used to optimize their inventory to ensure they have just the right amount of products to meet customer demand throughout the fiscal year. The company can also use their data to understand and predict customer and consumer preferences. For small, local businesses, these problems are easy to deal with. However, for large, international chains and big-box stores, where millions of transactions are regularly generated (sometimes in a single day), the problem presents a computational challenge where data mining methods are designed to excel. Finally, the mere fact that most retailers store every transaction you make and use these data for targeted advertising among other things, presents privacy concerns.

3.3.1 Optional, advanced topics

There are a wide range of topics that may or may not be included in an introductory course in data mining, depending on the course design goals, and the audience background. You might consider going into some advanced techniques that provide some depth on topics covered in the course. For example, some useful topics for depth that you might want to consider are:

- Probabilistic methods for data mining, such as Bayesian networks
- Advanced classification techniques, such as the use of artificial neural networks, support vector machines, random trees, random forests, or ensemble classifiers, which are methods that combines individual classifiers to generate one classification.
- Advanced clustering techniques, such as graph-based clustering methods

Alternatively, you can increase the breadth of topics in data mining by considering the addition of topics beyond the core topics suggested above. These topics include, but are not limited to:

- Data warehousing and OLAP for data mining
- Mining time series and other sequential data
- Data visualization methods for data mining

4. Tools and Techniques

Different types of data mining tools are available in the marketplace, each with their own strengths and weaknesses[13]. We recommend traditional data mining tools, which contain a wide range of methods for preprocessing data and performing common data mining tasks. Fortunately, there are numerous tools available; many are open-source and free for academic purposes.

A good exercise should motivate the students toward learning and using the tools to generate some good results. For simplicity purposes, we will use Edgar Anderson's Iris dataset[14], which is perhaps the most widely used dataset for demonstrating data mining tasks, particularly those for classification problems. This classic dataset, which comes installed with a vast majority of tools available, has 3 classes of data, each of which represents a type of iris plant: Iris Setosa, Iris Versicolor, and Iris Virginica. Each class has 50 example instances in the data. Each instance has four attributes in addition to its class label: sepal length, sepal width, petal length, and petal width. The aim is to develop a model from the data that can predict the class of the flower, given the four measurements.

The tools we present here can be used to help the student analyze the data and develop a predictive model. For assignments, we require students to use a specific tool. However, for final projects, the students were allowed to pick the tool of their choice. We used Weka and R in our courses, two widely used, freely available tools for data mining.

4. 1 WEKA:

Weka is a collection of machine learning algorithms specifically selected and implemented for data mining tasks. It is freely available for download at http://www.cs.waikato.ac.nz/ml/weka/ . It is also worth noting that the authors of the software have a book published that extensively makes use of Weka[8]. The authors of Weka have implemented a wide range of methods for association and frequent pattern analysis, classification and clustering. The list of algorithms

implemented in Weka spans well beyond what any single course would be able to cover, and thus provides a good option for students to be able to explore each of the methods taught in any data mining course. Additionally, the vast majority of Weka can be used without a single line of coding required, making the tool ideal for those that wish to broaden the audience to include non-CS majors.

Though it has a command-line driven interface, it is far more useful and interesting for students to use Weka through its graphical user interface. We taught students to use Explorer mode through the GUI, which provides a good environment for stepping through a complete data mining task. It includes tools to open and visually explore your data, preprocess and filter data to reduce noise and eliminate outliers, and experiment with different data mining methods to observe the strengths and weaknesses of each. Additionally, there are visualization tools built into Weka that are good for visually observing distribution and density of your data. For those that are teaching tree-based methods, there is a decision tree visualization tool that allows students to view the complete tree induced by the method selected. Finally, Weka comes with several datasets that are ready for students to explore data mining.

Using the data visualization tools, you can quickly load in a file, and explore the distribution of each class with respect to each attribute. Without any filtering or preprocessing, it is easy to observe that petal length and petal width are both strong predictors of the type of flower (See Figure 1):

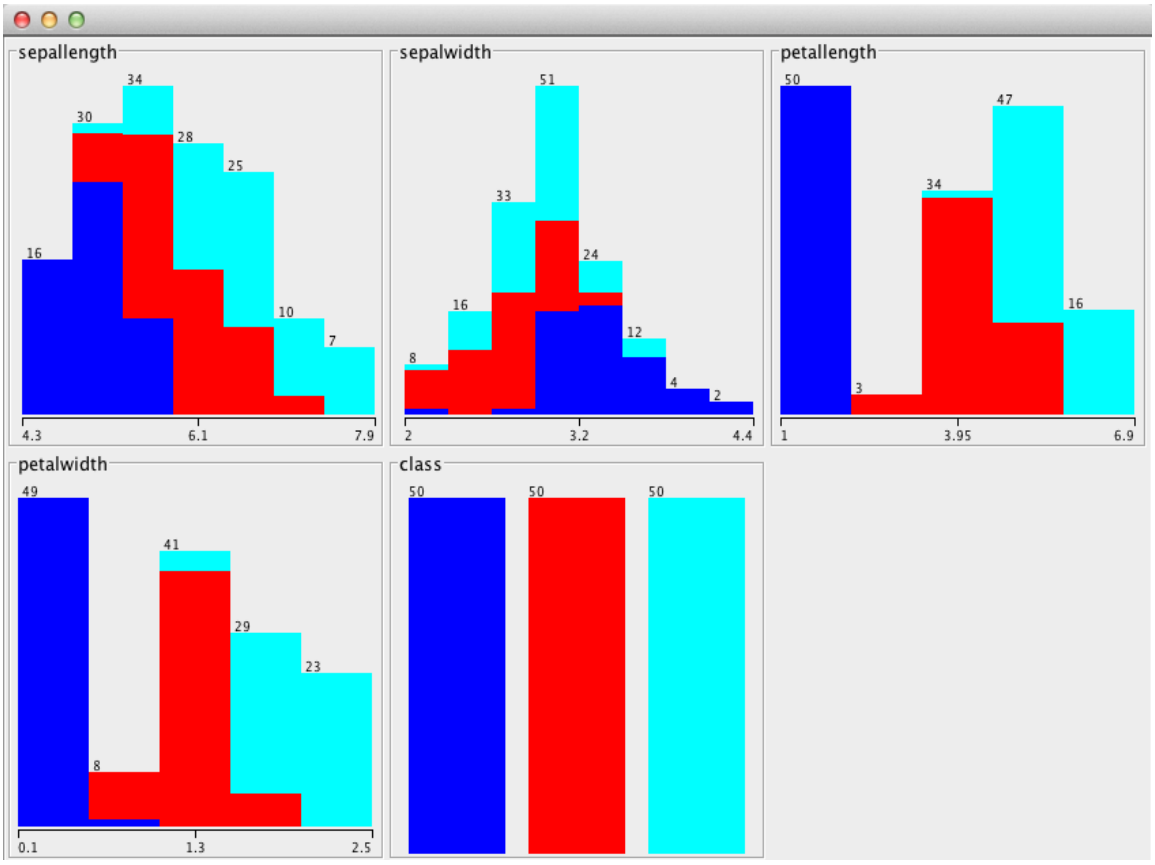

Figure 1 – Visualizing class distribution over many attributes in Weka

It has additional visualization tools that let you quickly observe scatterplots of all possible pairs of attributes in the data, giving you another easy way of observing which attributes most separate the data by class (See Figure 2):

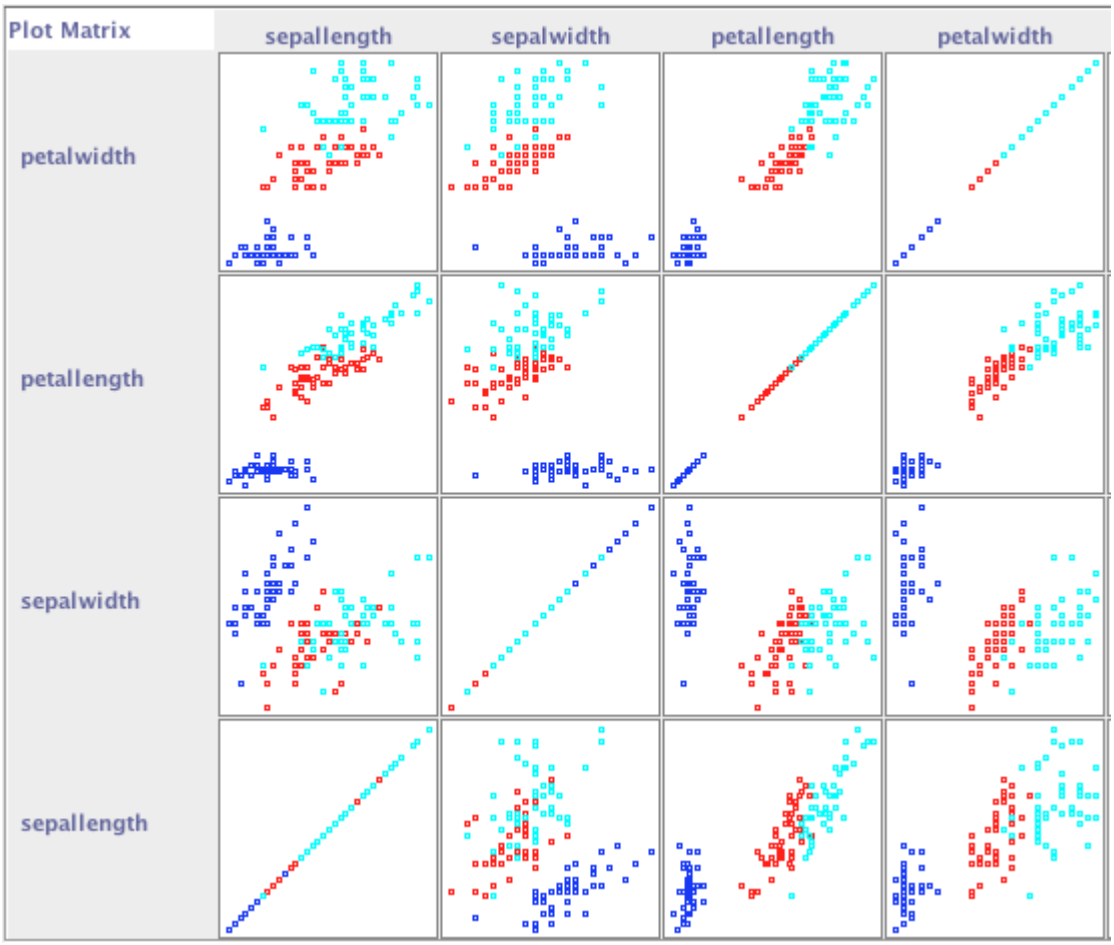

Figure 2 – Pairwise scatterplots of data, with points colored by class in Weka.

Once you are ready to build a classification model, Weka allows you to easily choose from a wide range of methods that induce a classification model. It also allows you to set up different validation techniques to ensure you are not testing your model against the training data -- the data used to build the model. All of the methods output useful information to help the student determine how good their method worked. For example, using the widely popular C4.5 decision tree algorithm[9] on the Iris dataset, the following output is generated, showing all of the common metrics for measuring classifier performance:

```
Correctly Classified Instances 144 96 %
Incorrectly Classified Instances 6 4 %
Kappa statistic 0.94
Mean absolute error 0.035
Root mean squared error 0.1586
Relative absolute error 7.8705 %
Root relative squared error 33.6353 %
Total Number of Instances 150
```

```
=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class
0.98 0 1 0.98 0.99 0.99 Iris-setosa
0.94 0.03 0.94 0.94 0.94 0.952 Iris-versicolor
0.96 0.03 0.941 0.96 0.95 0.961 Iris-virginica
Weighted Avg:
0.96 0.02 0.96 0.96 0.96 0.968

=== Confusion Matrix ===

a b c <-- classified as
49 1 0 | a = Iris-setosa
0 47 3 | b = Iris-versicolor
0 2 48 | c = Iris-virginica
```

Decision trees have an extra benefit in the Weka framework. The authors of Weka have provided a visualization tool to help the student understand what the induced decision tree looks like (See Figure 3):

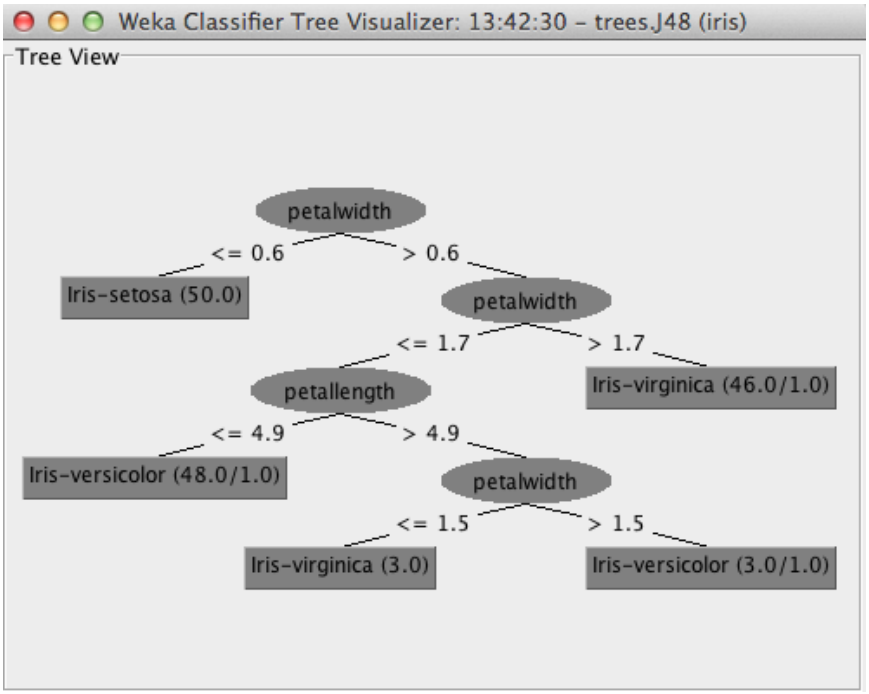

Figure 3 – Visualizing a decision tree in Weka

In a matter of minutes, you can easily demonstrate these examples right in the classroom. You can edit the data right within Weka if you wanted to demonstrate how anomalies can affect the final classifier. Additionally, you can compare the results of different models to see how different models or model parameters alter the output of the classification method.

For those with a bit bigger aspirations for their course, the Experimenter and KnowledgeFlow modes, which also are part of the GUI in Weka, are tools available to assist in larger, more

complex data mining tasks. They may be useful for student term projects. The Weka website has extensive documentation and tutorials for students to work through.

4.2 The R Project for Statistical Computing

R is a freely downloadable language and environment for statistical computing and graphics. You can obtain a version of R precompiled for a wide range of platforms from http://www.r-project.org/. The site has plenty of documentation and other helpful information to help get you started. R has an enormous number of capabilities owed in part to the large community of users worldwide that are actively developing new packages that extend the base distribution. The large set of available packages make this tool an excellent alternative to the existing (and expensive!) data mining tools. There is a good chance that if there is some task related to statistical analysis or data mining that you want to accomplish, someone else has already completed the work in R.

The base version of R comes with a simple environment that allows users to enter commands interactively or run existing scripts. However, to do anything useful requires learning the R language. We provided a couple of assignments for our students that required them to read through the R documentation and answer some questions that required them to write some very short commands and functions.

For example, suppose you wanted to complete a simple data visualization and classification exercise using the same Iris dataset demonstrated above. You would first need to obtain the data. There are a wide range of packages available that contain datasets, one of which is the "datasets" package. This package is included with the base R installation, but just incase, you can execute the following commands to install the library and load it into the environment:

```
> install.packages("datasets")
> library("datasets")
```

One of R's greatest strengths is its graphical output capability. For example, to view a matrix of scatterplots, you can use the `pairs` function, which is part of the "graphics" package, also included with the base installation (See Figure 4):

```
> pairs(iris[1:4], main = "Anderson's Iris Data", pch=21, bg =
c("red", "green3", "blue")[unclass(iris$Species)])
```

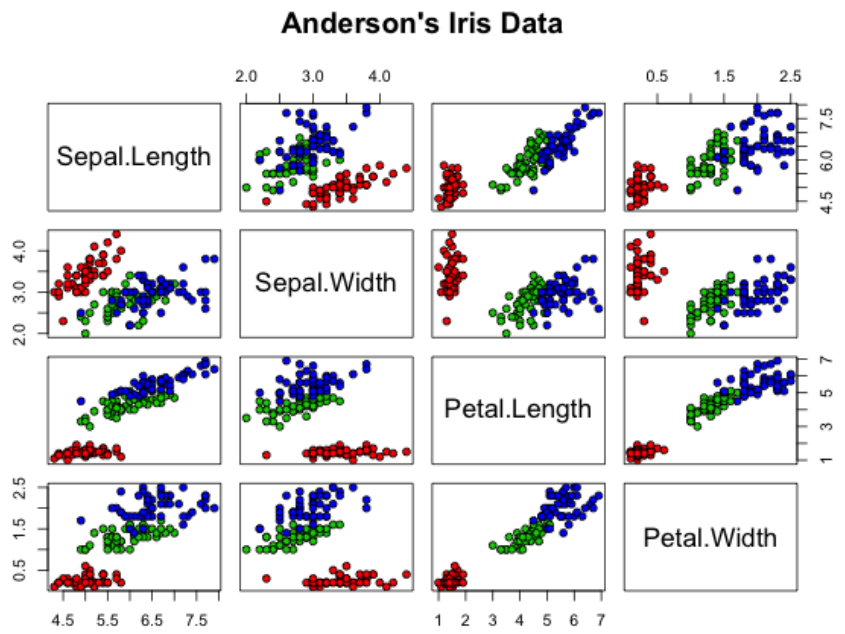

Figure 4 – Generating a scatterplot matrix in R

To induce a decision tree, you need to install the `rpart` package. The commands below will install the package, induce the classification model, and plot the tree:

```
> install.packages("rpart")
> library("rpart")
> fit <- rpart(Species ~ Sepal.Length + Sepal.Width + Petal.Length +
Petal.Width, data = iris)
> par(mfrow = c(1,2), xpd = NA)
> plot(fit)
> text(fit, use.n = TRUE)
```

The following tree is output (See Figure 5):

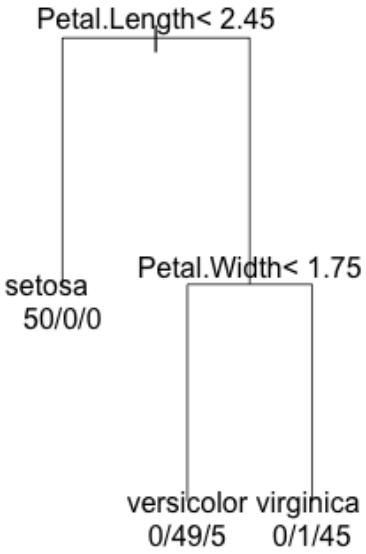

Figure 5 – A pruned decision tree visualized in R using the rpart package.

There is no doubt that, if you have never used R before, the syntax and function calls above look daunting. Indeed, this is the most intimidating part for students learning R. Learning the syntax of the language is only half of the battle. More frustrating is the process of discovering the wide range of packages and included functions available. Once you see a function that does what you want, it often has an enormous number of parameters to specify. For many basic tasks, there are often multiple paths to achieve whatever you want to do. Students expressed difficulty in trying to figure out which path was the best. We expected students to learn these aspects of R by using the web to their advantage; most were able to find a small handful of good websites for reference purposes and were able to conquer some of the initial intimidation that set in.

Perhaps the most important R commands to teach students are the commands for getting some help. Our first set of exercises are designed give students the facilities to confidently use the built-in help facility:

`> library(help=rpart)`	Display help on all commands in "rpart" library
`> help(rpart)`	Display help on the "rpart" command
`> ?rpart`	Same as above
`> help.search("classification")`	Search the help system for the word "classification"
`> example(rpart)`	Auto run through examples for the rpart command
`> vignette()`	Display a list of all available vignettes
`> vignette("longintro")`	Display the vignette called "longintro"
`> RSiteSearch("decision tree")`	Search the R Project pages for "decision tree"

Note that the `RSiteSearch` command will automatically open your browser window and execute the search on the R site using the specified search string. You can fine tune your search on the page that is brought up, as it is quite likely that hundreds of hits will be returned. However, if you are not sure where to begin with a particular task you are trying to achieve, this command is invaluable.

4.2.1 Websites for learning R

Using your favorite search engine will likely reveal hundreds of web sites for learning R. In addition to the main R project web site, http://www.r-project.org/ , we have a few web sites that we found students using the most for their learning:

- Quick-R - http://www.statmethods.net/ - This site is authored by Robert Kabacoff, a professional statistical consultant and research methodologist. It contains a wide range of examples that are presented in way that is very easy to understand, and easy to apply in your own work.
- RDataMining.com - http://www.rdatamining.com/ - This website presents numerous resources on using R for data mining tasks. As of this writing, there is a complete PDF version of the author's book, titled, R and Data Mining: Examples and Case Studies, to be published soon. Numerous examples and exercises from this material can be used as a basis for your own exercises and in-class demonstrations
- Data Mining Algorithms in R - http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R - This is another source of material for students to learn how to use R for basic data mining tasks.
- R Graph Gallery - http://gallery.r-enthusiasts.com/ - We had a few students that were intrigued by the excellent graphing capabilities of the R environment. This site has some wonderful examples of some complex graphs, with complete R code included for students to try out on their own

4.2.2 - Rattle - A Data mining GUI for R

We encouraged our students to work through the R exercises to learn the framework. Ideally, they would leave the course knowing R well enough to comfortably explore new data and apply basic data mining techniques to them. However, if your audience consists of a large number of non-programmers, R will be intimidating for them. If you want the power and flexibility of R, with the GUI friendliness offered by Weka, there is an optional package you can install in R called Rattle[10]. Rattle provides an easy to use entry into sophisticated data mining using R. Like R, it is free, open source, and runs on all major platforms.

To download and install the Rattle framework, from the R prompt, execute the following commands:

```
> install.packages("rattle")
> library("rattle")
> rattle()
```

Rattle uses a large number of existing libraries developed by the user community, and ties all of them together in a single GUI. For example, you can set up the same Iris dataset for analysis right from the GUI without doing any coding (see Figure 6):

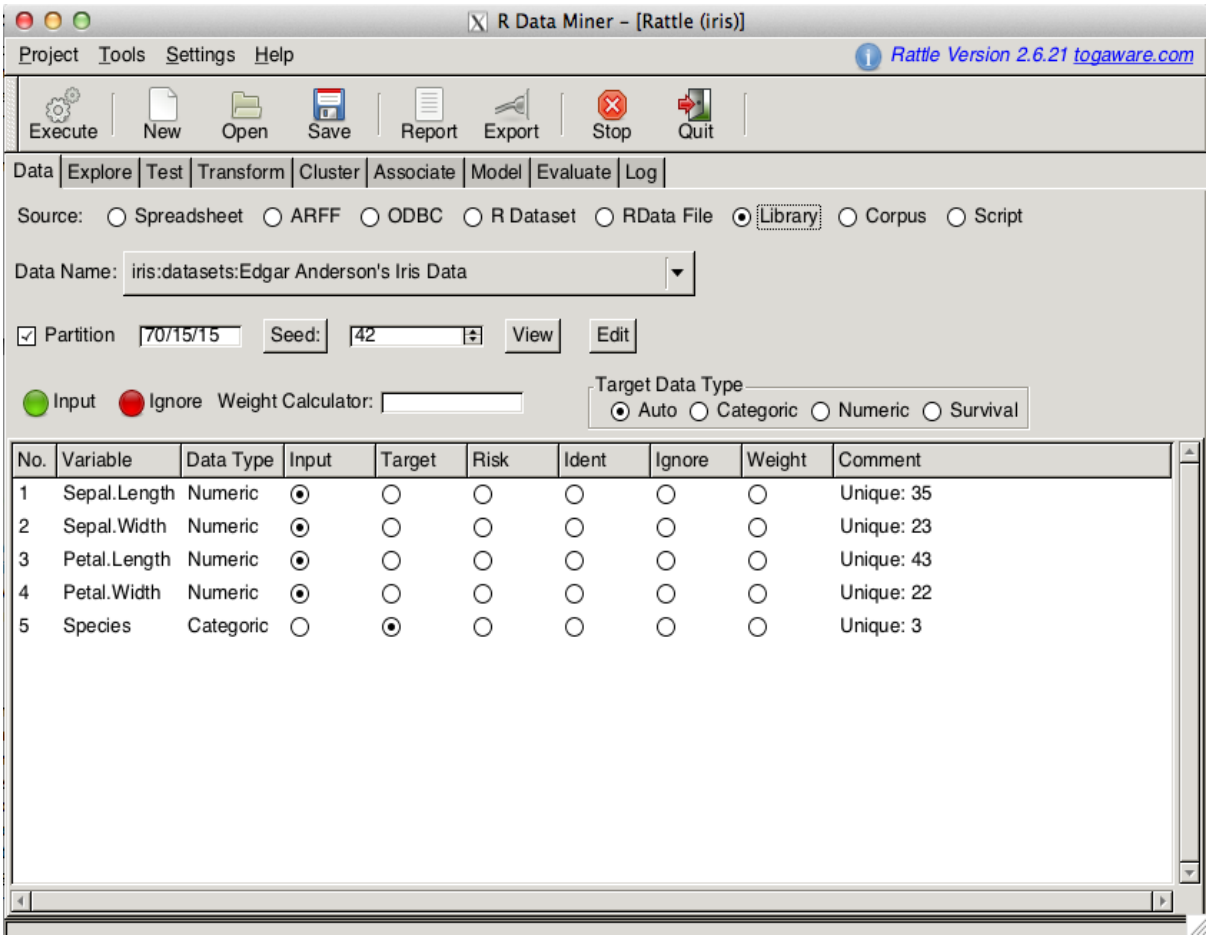

Figure 6 – The Rattle GUI for data mining in R

Rattle contains a small number of methods for association and frequent pattern mining, classification, and clustering. Though the number of methods included is far fewer than presented in the Weka framework, it contains everything needed for an introductory data mining course. As the author's web site states, "The aim is to provide an intuitive interface that takes you through the basic steps of data mining, as well as illustrating the R code that is used to achieve this. Though the tool itself may be sufficient for all of a user's needs, it also provides a stepping stone to more sophisticated processing and modeling in R itself, for sophisticated and unconstrained data mining." (See http://rattle.togaware.com for more information.)

4.3 Other frameworks

There are increasing options available for data mining software frameworks. We covered Weka and R, mostly because these are the environments we choose for our course. Both of these are widely used in the user and research community. R particularly scales well toward Big Data. However, our students did indicate that the learning curve for R was steep. Computer science students, as well as students that knew Matlab or Python prior to taking the course had the easiest time learning the language. Depending on the background of your audience, there are alternative options that might have an easier learning curve.

- Python has been quite popular in the scientific community, with the NumPy and SciPy packages being the most widely used packages for numerical and data analysis. Additionally, Python is starting to have some decent data mining and data visualization tools available. One Python package worth noting is Orange (http://orange.biolab.si/), which provides an excellent visual programming interface, allowing a large number of basic data mining tasks to be completed without any Python coding[15]. Another useful package is scikit-learn (http://scikit-learn.org), which is a collection of useful machine learning algorithms for analysis and visualization built around the popular numpy and scipy modules in Python[16].
- Matlab is a very popular commercial framework for numerical computing. It is a popular system in the academic world, particularly in colleges and universities with engineering programs. Like R, it includes an environment for interactive sessions, as well as a complete language for writing scripts. For data mining purposes, the most useful toolbox is the Statistics Toolbox (http://www.mathworks.com/products/statistics/). Matlab also has an active user community with a wide range of scripts and complete toolboxes available that are written to accomplish the majority of data mining tasks.

5. Data

Despite the real-world application of data mining to big data, for pedagogic purposes, the vast majority of topics in data mining are taught using small data sets. This makes it easy for students to easily work through many of the algorithms that are used with the methods. Weka and R both provide a wide range of datasets available for learning purposes:

- Weka: All data sets are installed in the directory where Weka is located. Weka uses its own Attribute-Relation File Format, or ARFF. Weka has tools available that can read and / or convert data in other popular formats.
- R: The "datasets" package is the place where you will find the majority of the data you can use for basic examples.

For advanced projects and assignments, you will want to provide sources where students can easily get their hands on real-world data. In fact, the most common criticism we heard at the end of the course was the lack of assignments that gave them an opportunity to see how the methods being taught would scale on large datasets. Below is a list of sites that contain some relevant, real-world, large datasets that will give your students places to consider finding data for their own project:

- The UCI Machine Learning repository – http://archive.ics.uci.edu/ml/index.html – One of the most well known and widely used sources of data for research purposes. It currently hosts 235 different publically available datasets (as of this writing). The majority of the students found data to use for their project here.
- Kaggle – http://www.kaggle.com/ – Kaggle is an increasingly popular website which houses a large number of datasets from current and former data mining competitions. These competitions represent opportunities for students to take part in real-world data mining competitions using data contributed by a wide range of industries that are hoping to glean new knowledge from data. Many of the competitions offer hundreds, thousands, or even millions of dollars in prizes. It is worth noting that nearly all of the students took

the time to explore the current competitions, and nearly half of them set up an account to download the data and see how tough it might be. We had one student that was determined to win the "Don't Get Kicked" competition held in Fall, 2011 (http://www.kaggle.com/c/DontGetKicked). In this competition, contestants were competing for a $10,000 prize to develop the best model to predict the whether a used car for sale at a auction is a good buy or not. His submission placed in the middle of the pack of over 500 teams worldwide, many of which consisted of teams of graduate students and retired professors – a respectable finish for an undergraduate student who just learned basic data mining.

- ACM KDD Cup Archive – http://www.sigkdd.org/kddcup/ - KDD Cup is the annual competition hosted by ACM SIGKDD – the most widely known professional organization of data miners. With ACM being the premier association for computer scientists worldwide, we encouraged students (especially computer science students) to visit this site.
- Data.gov – http://www.data.gov – The US Government is working hard to continually make its data publicly available. A wide range of datasets are continually being made available, covering categories including agriculture, marriages, deaths, health, banking, elections, environment, education, and military, to name a few.

There are many other sites to choose from, and stud

6. Discussion

Overall, we were pleased with the results of the first offering of this new course. We took a great deal of time to review an enormous amount of material, both in print and online. We surveyed the students by requiring them to make an entry in a private, online journal once per week, where they were free to express their thoughts about their experience in the course. In this final section, we first present a general assessment of our learning outcomes through offering data in our final project. Then, we present recurring themes from our own observations and from the student journal entries that will hopefully help the reader form a new course with "eyes wide open."

6.1 Assessment of Learning Outcomes

To assess the learning outcomes of the course, students were required to find a real-world problem where data mining could be used to provide useful knowledge about the underlying processes that generated the data in the first place. Specifically, they had to find the data to use on their own, understand and explain the data and where they came from, apply appropriate preprocessing techniques, evaluate a minimum of three different methods using standard performance criteria taught in class, and most importantly, give a proper interpretation of their results. The project culminated in a final paper and an in-class presentation. They were required to work in teams of 2-3 students each. The project was extensive, requiring a 2-page proposal explaining what data they will mine, why it's important, and what they might hope to learn (10%), write a paper giving the reader an appropriate background on the data, the methods applied, a comparative assessment of the results obtained, and a discussion of their reuslts (70%), and a 10 minute in-class presentation of their findings (20%). The rubric for the final project focused on important facets of any small, in-class research project. An important part of the

rubric was weighted toward those outcomes that were most important for the class – proper use of existing classification methods, and proper interpretation and evaluation of the results obtained. These were evaluated on both paper and in presentation. Because of the wide range of programming strengths in the class, we allowed the students to choose pure GUI-based data mining tools, use R with existing packages, or implement their own system using their favorite programming language.

As a first-time offering of the course, we were admittedly a bit more lenient with the grading than we would have liked. However, we were quite pleased with the results for the first offering of the course, and in some cases, pleasantly surprised. The grades ranged from 77-99, with the average grade of 90 observed.

The vast majority of low performing teams suffered for the usual reasons undergraduate teams do poorly on final projects – procrastination. The lowest performing team was a combination of procrastination and poor understanding of the tools due to lack of quality time spent on preceding homework assignments. We noted that the students that had less background in computer science were far less likely to do any programming, instead opting for the Weka GUI-based software.

We found that the vast majority was quite pleased to work on an interesting, relevant large-scale dataset of their choice, and see how the methods taught would work in practice; their enthusiasm was reflected in the results obtained.

6.2 Reflection and Discussion

One of the biggest challenges we faced with the design of the course was from the unexpected interest from non-CS majors. While this was a pleasing observation, it did require us to reconsider the depth of some material, and perhaps consider some different techniques in the future, as the interest is continuing to expand. We are strongly considering offering two variants of the course: one course would be the existing data mining course as an elective for the computer science major, with a prerequisite of taking a course on data structures, algorithms, and a statistics and probability course. An alternative course with an interdisciplinary focus is being discussed that would be co-taught by multiple individuals from different departments. This would be offered as an introductory course, where very little programming would need to be completed, with the vast majority of labs, homework and projects coming from written exercises with small, simple example datasets, or require the student to use the graphical-based Weka data mining framework. If we indeed split the course, then we would make our elective course a bit more advanced by requiring our algorithms course as well, and including more advanced projects where programming would be required. This topic is still under discussion.

Regarding prerequisites for the course, we recommend that course designers require a course in statistics and / or probability as a definite prerequisite, preferably completing the course with a reasonable minimum grade. For our first offering of the course, we were more relaxed about this prerequisite than we should have been in order to allow a broad audience. Fortunately, only two students did not have the desired level of statistics background, and their understanding of some material on understanding distributions of data, skew, correlations, and related concepts was limited. We encourage you to resist the temptation to lower this standard.

Another challenge we faced was developing short labs and assignments that gave students an opportunity to use real-world data. For obvious reasons, the vast majority of exercises for labs and assignments in data mining courses use very small datasets. While this is important for teaching the different methods in data mining, it prevents students from experiencing the real reason data mining exists in the first place. We witnessed some students expressing frustration toward the middle of the course because they wanted to see how the methods we were teaching worked on datasets that can be classified as Big Data. This concern was alleviated when students were given their final project, where they were required to use a large dataset for analysis. An additional assignment or two should be given well before the final project that uses a large dataset for analysis.

The ubiquitous nature of data should leave no excuse for students to fail to find an interesting dataset to work with. In fact, about a third of the projects that were submitted by students came from data they found on their own, outside of any of our suggested sites. For example, we had a team that worked with the admissions department on campus to collect some data in order to develop a model that can predict the students most likely to succeed at our campus. Another team set up a Twitter account to develop an application to apply constrained clustering techniques on large collections of tweets to determine sentiment within the clusters associated with a publically traded stock. The comments from these teams suggested they liked being able to form their own small study with their data.

As mentioned before, a large number of students complained about the learning curve associated with R. For our first offering of the course, we put a lot of the burden of learning R on the students, pointing them to the documentation on the web site. We expected students to work through the exercises on their own, with only minimal credit offered for completing a few exercises. In hindsight, this was a mistake. To reduce the initial intimidation and frustration with learning R, we cannot emphasize enough how important it is to introduce R with simple exercises in a tutorial fashion, preferably with a lab setting. This helps students be able to focus solely on learning R. The first lab should simply focus on the syntax of the language, with subsequent exercises designed to incorporate aspects of R's powerful data analysis library. On a related note, we did not offer this course as a lab course. When surveyed at the end of the semester, students did not seem to think the course should be a lab course, however, they did state that learning R would have been easier for them if a few classes were set aside solely for learning R. For the next offering of the course, we have set aside two weeks of the semester to move into the lab and have them work through numerous exercises designed to get them learning the R framework. We highly recommend one of the integrated environments that run on top of R, such as RStudio (http://www.rstudio.com/). Our students largely favored RStudio over the default R environment, and widely agreed that this improved their experience with learning and using R.

A common misconception is that the world of Big Data is mostly of interest only to students with some involvement in information technology. We encourage you to dispel this myth with your students very early in the semester. For example, ask your students to consider the non-technical implications of data mining and Big Data. The most distressing topic that many of us think of with data mining revolves around issues of privacy and security of the data being mined. Consider the legal issues surrounding the intellectual property and ownership rights for protecting not only the data itself, but more importantly, the interesting patterns that are allowing

companies to make some big money from their data. Their data consists of records that often represent things that they do not own. As businesses are realizing the potential pool of wealth hidden in their data, they are also realizing the technical *and* non-technical infrastructure that needs to be put in place to enable them to use their data in a way that brings them the hidden knowledge without hindrances.

We note that some colleagues have developed data mining courses that tend to focus on reading through current literature in data mining, rather than use any single textbook. Musicant noted that his approach toward using a combination of current literature and programming assignments worked well[17]. We may incorporate some current readings from data mining literature into the course in the future. Undergraduates often have very little experience with reading current papers beyond a textbook before they graduate. Data mining and Big Data have a plethora of literature to read; this would be a good addition to consider.

7. Conclusion

Much of the work presented here is based on our collective research and industrial experience with data mining, which culminated in the authors strong interest and desire to teach data mining to a new generation of undergraduate students in STEM fields. Students that are learning in the era of Big Data are observing that nearly every aspect of their world around them has data being collected. Their computers, cars, appliances, mobile devices, gaming consoles, and shops they visit all are collecting data with every activity they perform. The vast majority of industries that our students are seeking careers in have a plethora of opportunities to use data mining to further their potential employers.

Traditionally, data mining has been widely associated with programs in computer science. It is our hope that you are able bring the world of Big Data to a wide range of disciplines. We presented a variety of tools and topics that can help you bring this important topic at a level that makes sense for your students.

We are doing a disservice to our students by not bringing the topic of Big Data to the forefront of their education. Now is the time for educators in all STEM fields to work hard to provide our students with the knowledge required to competently work in the emerging field of Big Data.

8. References

[1] James, J. "How Much Data is Created Every Minute?" Retrieved November 3, 2012 from http://www.domo.com/blog/2012/06/how-much-data-is-created-every-minute/?dkw=socf3

[2] Gartner Press Release. "Gartner Says Big Data Will Drive $28 Billion of IT Spending in 2012." October 17, 2012. http://www.gartner.com/it/page.jsp?id=2200815

[3] Gartner Press Release. "Gartner Says Big Data Creates Big Jobs: 4.4 Million IT Jobs Globally to Support Big Data by 2015." October 22, 2012. http://www.gartner.com/it/page.jsp?id=2207915

[4] Manyika J, Chui M, Brown B, Bughin J, et al. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute Report. 2011.

[5] IBM Corporation. "What is big data?" retrieved November 2, 2012 from http://www-01.ibm.com/software/data/bigdata/

[6] Han J., Kamber M, and Pei, J. Data Mining, Concepts and Techniques, 3rd ed. Morgan Kaufmann. 2012

[7] Chakrabarti, S. et al., Data mining curriculum: A proposal (Version 1.0). Intensive Working Group of ACM SIGKDD Curriculum Committee, April 2006, retrieved October 15, 2012 from http://www.sigkdd.org/curriculum/CURMay06.pdf
[8] Witten I, Frank E, Hall M. Data Mining: Practice Machine Learning Tools and Techniques, 3rd ed. Morgan Kaufmann, 2011.
[9] Quinlan, R. C4.5. Programs for Machine Learning. Morgan Kaufmann Publishers, 1993
[10] Williams, G. Rattle: A Data Mining GUI for R, Graham J Williams, The R Journal, 1(2):45-55. 2009.
[11] Torgo, L. Data Mining with R: Learning with Case Studies, Taylor & Francis. 2010.
[12] Aldana, WA. Data Mining Industry: Emerging Trends and New Opportunities. MIT Dept. of Electrical Engineering and Computer Science. MIT Press, 2000.
[13] Silltow, J. Data Mining 101: Tools and Techniques, posted August 2006, retrieved December 1, 2012 from http://www.theiia.org/intAuditor/itaudit/archives/2006/august/data-mining-101-tools-and-techniques/
[14] Fisher, RA. The use of multiple measurements in taxonomic problems. Annnual Eugenics, 7, Part II, 179-199, 1936.
[15] Curk T, et al. Microarray data mining with visual programming. Bioinformatics 21(3):396-8, 2005.
[16] Pedregosa et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research (12):2825-2830. 2011.
[17] Musicant DR. A data minig course for computer science: primary sources and implementations. SIGCSE '06 – Proceedings of the 37th SIGCSE technical symposium on computer science education. 538-542, 2006.