Data Mining Methods Course for Computer Information Systems Students

Musa Jafar
mjafar@mail.wtamu.edu

Russell Anderson
randerson@mail.wtamu.edu

Amjad Abdullat
aabdullat@mail.wtamu.edu

CIS Department,
West Texas A&M University
Canyon, TX 79018

Abstract

Although a Data Mining Methods course sequence is a late comer to the Information Systems curriculum, it is a natural fit in the discipline and students largely benefit from it. Students graduating with a BBA degree are well prepared for this area of specialization. They do understand business processes (approximately forty percent of their course work is in the business discipline and in quantitative analysis). Business knowledge coupled with knowledge of computing, and data management uniquely prepares those students for excellence in the discipline. In this paper, we present the implementation of a junior-senior level elective data mining methods course that we designed and offer as part of our BBA in Computer Information Systems. The course is business oriented. It emphasizes the conceptual understanding of data mining theory, practices and the current state of the underlying computing technology. We use off-the-shelf tools for the homework assignments and projects to perform the data mining tasks (cluster analysis, association analysis, decision tree analysis, naïve Bayes, neural network, etc.). In this paper, we also present the supporting technologies and resources that we used for the course, along with lessons learned from teaching the course.

Keywords: Information Systems, Data Mining course, Business Intelligence, Data Modeling, Business Processes.

1. INTRODUCTION

In 2007, IBM paid 5 billion dollars for Cognos, SAP paid 6.8 billion dollars for Business Objects, and Oracle paid 3.3 billion dollars for Hyperion (Pendse 2007). The acquiring companies are among the leading five software and services companies on the Forbes 2007 list. The acquired companies' core competencies are focused in the area of business intelligence solutions. These acquisitions are significant because they clearly indicate that the leaders in the software and services industry are heading towards non-traditional analysis of large data sets for the purpose of decision support. Microsoft's name is not on this list because they relied

on internal development to strengthen their market presence in the area.

"The field of data mining grew out of the limitations of current data analysis" (Tan 2006). The advancements in machine learning algorithms, pattern recognition algorithms, and artificial intelligence at large coupled with current trends in computing (CPU power, massive storage devices, connectivity and affordability) enabled universities to bring the data mining course curriculum content, theory and supporting technology into the classroom. Before 2000, it was a challenge for an academic department to realistically and consistently incorporate the theory, practices and the technology into the curriculum at a level where the course content are well defined and the underlying technology is affordable, at an acceptable level of performance and stable enough without having students write the algorithms themselves instead of using off the shelf stable products (Lu, 2002). The maturity of data mining algorithms and the advancement and affordability of computing machinery have made it possible to uniformly offer courses in the discipline at the undergraduate level and allowed researchers to even explore alternative learning styles for the offerings of such courses (North 2007).

Lenox (2002) surveyed the different offerings of undergraduate courses in data mining. They found that, "Although data mining is of increasing interest to industry, it does not appear to be a common undergraduate course at the time". However, in the past six years this trend has dramatically changed. Currently, data mining methods courses are offered at the graduate and undergraduate levels, in CIS, MIS, CS, Statistics, and Informatics programs under titles such as Data mining, Business Intelligence, Knowledge Discovery, Cyber Security, Advances in Databases, Machine Learning, Artificial Intelligence, Bioinformatics, Advanced GIS, etc. The contents of the course have matured to the level that there is not much difference in content between the course offerings. In their papers (Lu 2002, Lenox 2002, Roiger 2005, Musicant 2006, Saquer 2007), made it clear that the content of the course and the underlying technologies have evolved from an experimental course content with heavy emphasis on the coding of the algorithms to a main stream course where content is well defined and the emphasis is on using stable and off the shelf data mining tools.

2. WHAT IS DATA MINING

Data Mining has been defined as "the process of discovering useful information in large data repositories" (Tan, 2006) for the purpose of decision support. It can also be thought of as the integration of quantitative analysis, database management, data visualization and intelligent computing (machine learning, pattern recognition and artificial intelligence) for the purpose of discovering patterns that are not previously known in the data.

Data mining analysis does not happen in a vacuum, the analyst needs to understand the underlying business processes, the characteristics of the data set within the context of the business, and the conceptual model of the underlying data. Analysts need to be clever and have insight capabilities into how to analyze data within context. They need to understand the characteristics of each attribute, to a certain extent the key influencers of each attribute, and be able to aggregate, consolidate, extract and transform data.

The data mining process has been defined as an Extract, Transform and Load (ETL) process to consolidate data from different sources into a unified and consistent repository. This process assumes that all of the preprocessing, scrubbing and consolidation of data has been performed. However, in a corporate environment, that is not the case. Accordingly, a pre-ETL process needs to exist. The process starts with a data set that may: (1) need to be understood within the context of a business environment where the characteristics of the data and the business rules of the environment need to be understood; (2) need to be cleaned-up or scrubbed, aggregated and consolidated; and (3) need a unified logical schema to be defined. After this the formal extract, transform and load process in the traditional sense of data mining is applied. So, a data mining analyst is a clever person with strong data modeling, business processes, quantitative analysis and programming skills, with a conceptual understanding of data mining theory and a comprehensive understanding of data mining tools, techniques and practices. Although it is a plus, an analyst does not need to in-depth knowledge of

the theory of Support Vector Machines, Neural Network computing, etc. to the level where they can dissect and write the code for algorithms to perform data mining analysis tasks.

Section three of this paper is a lengthy section. It provides the details of the course objective, contents and organization. Section four provides references to available data sets to be used in a data mining course. Section five provides references to learning resources available for a data mining course too. Section six provides our summaries, conclusions and lessons learned.

3. COURSE CONTENT

Basically a standard data mining textbook covers (1) classification algorithms such as decision trees, Bayesian-based algorithms, artificial neural networks, support vector machines, etc.; (2) association analysis algorithms which is also known as market basket analysis; (3) cluster analysis algorithms such as hierarchical, graph-based, center-based, etc.; and (4) statistical analysis and anomaly detection. For a textbook, we used "Introduction to Data Mining" by Tan, Steinbach and Kumar from Addison Wesley (Tan 2006). It is a standard data mining text for computer science students. We heavily augmented the content of the textbook to soften the course material for a BBA student. Two other books that we found helpful are "Data Mining Concepts and Techniques" (Han and Kamber 2005) and "Data Mining Practical Machine Learning and Techniques" (Witten and Frank 2005). Another book "Data Mining with SQL Server 2005" (Tang 2005) was valuable in supporting the Microsoft Business Intelligence environment. For future offerings of the course, we will seriously consider adopting (Han and Kamber 2005) and use (Tang 2005) as a reference. Our opinion is also supported by (Letsche 2007) where they surveyed a list of potential candidates for an undergraduate data mining course. They did not include (Tan 2006) in the list of books they compared. KDnuggets.com published a Poll of 200 hundred respondents to rate Data Mining Textbooks. The Han and Kamber text was rated top with 24% of the votes. (Lenox 2002) also provided a list of data mining textbooks that are consistent with Letsche (2007) and KDnuggets.com views.

The course we offer is a three-credit hours, junior-senior level. It is an elective course for students majoring in Computer Information Systems. The syllabus states that, "Upon the successful completion of the course, students should understand data mining concepts, and should be able to analyze large sets of data using data analysis tools for the purpose of decision support". As a prerequisite, the course requires knowledge of business statistics, database management and introductory programming. Statistical analysis knowledge allows students to understand the role of statistical analysis in data mining algorithms, perform statistical analyses, generate and interpret descriptive statistics, find outliers, plot histograms and box plots tasks. The database management knowledge allows students to understand the role of data modeling, data representation, data retrieval and data manipulation. It also allows them to build database schemas and load data into the backend database engine. The programming knowledge allows students to write scripting code for the purpose of data preprocessing cleanup, conversion, aggregation, normalization and consolidation; it provides them with the knowledge needed to build mining models using business intelligence tools.

The objectives of the course are consistent with the seven objectives stated in Lenox (2002) based on the "ACM Computing Curricula 2001 Ironman Report". We find these objectives still valid and serve as a benchmark for an undergraduate junior-senior level course like ours. We added two more objectives: eight and nine. Our revised list follows, our additions and comments are underlined:

1. Given several data sets, students will be able to understand and discuss ways to prepare (clean) and warehouse data. Students should be able to use a scripting language to clean the data sets.

2. Given a prepared set of data, students will be able to classify data based on association analysis, cluster analysis, decision trees, Bayesian networks or backpropagation algorithms. Students should be able to explain the accuracy of the algorithm and compare the performance of the various algorithms for the same mining model.

3. Given a prepared set of data, students will be able to analyze the data set using one of the techniques discussed.

4. Given an analysis of a particular data set, students will be able to make appropriate predictions.

5. Students will be able to understand, adapt and make proper selection of algorithms.

6. Students will be able to outline the applications of data mining techniques.

7. Some students could be motivated to continue Computer Science research or graduate studies.

8. Given a data set, students should be able to use statistical analysis tools to identify outliers, plot histograms, identify correlations, etc.

9. Given a data set, students should be able to use off-the-shelf visualization tools to analyze the data.

We identified six major relevant topics and activities to be covered in support of the preceding nine objectives as follows:

1. Data types, data quality, data preprocessing, summary statistics, data visualization and outlier detection. (chapters one, two, three and ten of the textbook).

2. Various similarity measures with emphasis on the concept of entropy (chapter two of the textbook).

3. Clustering and clustering algorithms with emphasis on center based clustering and density based clustering (chapter eight of the textbook).

4. Association Analysis with emphasis on market basket analysis and the underlying algorithms (chapter six of the textbook).

5. Classification Analysis including Decision Trees, Naïve Bayes and Artificial Neural Networks concepts with emphasis on decision trees algorithms and information gain, impurity measures, training models, testing models, prediction models, lift charts and comparison of performance among the classification models (chapters four and five of the text book).

6. Tutorials and demos on how to use an off-the-shelf data mining tool set in support of the course. (We do emphasize the importance of an off-the-shelf data mining tool set, for an undergraduate BBA course).

Course Organization-Learning Environments

The course was organized around a Presentation-Practice-Production-Evaluation paradigm. **P**resentation of the concepts through lectures, **P**ractice of concepts through hands-on tutorials using several software tools, **P**roduction of homework assignments and the final project and **E**valuation through exams, quality of assignments and final project. The course was conducted in a computer laboratory classroom. Although we emphasized a hands-on approach using specific tools, it should not be misconstrued as a training course in a particular software package. The course emphasizes the theory and practice of data mining. The computing resources that are available in the classroom include twenty-five PCs connected to a file server and the database server. The PCs are relatively new with above average specifications. We used off-the-shelf tools for the course and did not attempt to ask students to code any of the data mining algorithms. We do not think that this is necessary for an undergraduate course. We selected Microsoft's Business Intelligence environment for the course. Other than the bells and whistles and the internal implementation of the algorithms, there is not much difference between the various main stream data mining vendors' tools. They all use similar computational frameworks. For example, whether you use SAS Enterprise Miner, Oracle's ODM or Microsoft's Business Intelligence environment, they all support the basic supervised learning algorithms (Naïve Bayes, Decision Tress, Regression Models, etc.) and unsupervised learning algorithms (Association Analysis, Clustering, etc.). It is similar to teaching a database concepts class, where it does not really matter much if we are using Oracle, DB2, SQL Server or MySQL as the underlying technology. The Microsoft environment provided us with a stable, self contained (which is an important factor), reliable and easy to configure and use platform. Microsoft also provides a combination of tutorials, books on line and web casts that are easy to follow. The Microsoft environment also comes with a large database (Adventureworks Bike shop) that we use for homework assignments and for concept demon-

strations. This support provides us with an ideal rich environment at virtually no cost and minimum effort. The tools for the course included:

a) Microsoft SQL Server 2005 as a back end database management system and data mining engine. The environment is configured and maintained by the department with students given access and privileges to build databases and perform data mining tasks.

b) SQL Server 2005 Business Intelligence Development Studio as a front end desktop integrated development environment to manage data sources, create and manage business intelligence projects, mining models and mining model results analysis (parameter calibrations, Accuracy charts, predictions, comparisons, etc).

c) Microsoft Excel-2007 and SQL Server 2005 Data Mining Add-ins as an elementary data analysis tool. It is a good set of tools for students to dive in headfirst and start analyzing data. Microsoft also provides the large bike shop data set for the add-in and tutorials on how to perform basic data analysis tasks for the purpose of decision support.

d) In-house visualization tools built by faculty in the department for supporting courses and research. Figure 2 is a sample output of a parallel plot visualization tool that we used. Those tools support visual preliminary data analysis to gain insight into the characteristics of the data set and to visually analyze dependencies between attributes.

e) Excel, visual basic scripting and macros for data cleanup, statistical analysis, histograms and box plots and outlier analysis.

Homework Assignments

All of the homework assignments were in support of the six major topics presented in the course as stated earlier in this section as follows:

Statistical Data Analysis: Using Excel and Excel macros, students were required to generate 1,000 normally distributed grades between 40 and 100, compute summary statistics, quartiles, skewness, kurtosis, etc, and attempt to detect outliers. Compute the letter grades, plot their histograms, calculate the impurity measure of the grades (entropy, Gini index and classification error) and plot and compare the impurity measures for the purpose of gaining insight on the sensitivity of the three measures.

Data Clean-Up: Students were given a large data set with multiple Excel worksheets that maintained information about the toxicity level of chemicals as it relates to reproduction, growth and mortality of species. The data set contained more than 60% redundant entries, in addition to a lot of errors and inconsistencies. Students were required to write Excel macros to clean up the data and report errors and inconsistencies.

Visual Data Analysis: Students were given the Iris data set, a well known data set from UCI machine learning repository (UCI 2008). It contains records on the sepal width, sepal length, petal width, petal length and the classification of varieties of Irises (Setosa, Versicolor, or Virginica). For the purpose of comparison, understanding data ranges and outliers, students were asked to (1) perform various statistical analysis on the data, produce the overall box plots of each attribute at large and the box plot of each attribute within each iris variety individually, (2) develop a parallel plot of the data using a parallel plot tool that was developed in-house by the coauthors of the paper, the parallel plot tool allows for a simple interactive process (start application, load data, select attributes, view parallel plots, visually analyze the parallel plots, then visually manipulate the data on the plots). Figures 1 and 2 clearly depict the power of visual tools. Figure 1 is a display of the box plots collectively by individual attribute and by category within an attribute. Although the overall data set contains some outliers with respect to the individual attributes (AllSepalLength, AllSepalWidth, AllPetalLength, AllPetalWidth), when we break the data down by the iris variety category more outliers are visible. For example, in the top-right diagram (SepalLength comparisons), the AllSepalLength Box plot shows one outlier (diamond symbol, value 7.9), however, when SepalLength data was plotted within category (by iris variety) more outliers appear in the data for a total of 6 outliers (diamond symbols). This is only with respect to SepalLength attribute. The same argument applies to the rest of the plots in the figure. Students were able to recognize the importance of visual analysis

of data. For the same assignment, students were asked to use Parallel Plots to analyze the data. Figure 2 is a parallel plot of the iris data, each connected line is a data point (record), the top parallel line (variety) is the classifier and different iris varieties are displayed with different colors. The combination of box plots and parallel plots allow students to produce the following classification rules:

If PetalLength **is** VeryLow
and PetalWidth **is** VeryLow
and SeplaLength **is** Low to Medium
and SepalWidth **is** Medium to high
Then Iris Type **is** Setosa.

If PetalLength **is** Medium to High
and PetalWidth **is** Medium to High
and SeplaLength **is** Medium to High
and SepalWidth **is** Medium
Then Iris Type is Verginica.

If PetalLength **is** Medium
and PetalWidth is Medium
and SeplaLength **is** Medium to High
and SepalWidth **is** Low to Medium
Then Iris Type **is** VersiColor.

Market Basket Analysis: Students were given the SalesDetails, SalesOrderWholeSale and SalesOrderDetail data from the AdventureWorks bike shop data warehouse database. Using SQL Server 2005 Business Intelligence Development Studio they were asked to perform market basket analysis on the data to discover rules pertaining to what items are most likely to sell together and the corresponding rule support and confidence. Students were required to produce marketing reports on sales and promotion.

Clustering Analysis: Students were given the Iris data set and the Mushrooms data set from the UCI repository (UCI 2008). Using SQL Server 2005 Business Intelligence Development Studio they were asked to use different clustering techniques to find the characteristics of each cluster and the dependencies between clusters. The environment allows students to calibrate the algorithm parameters for a good fit of the number of clusters.

Prediction Analysis: This homework assignment was a lengthy one in preparation for their final project. Students were given the bike shop database. They were asked to break the database into a training set, a test set, and a prediction set. Then they were asked to build (1) a neural network model, a decision tree model, and a naiive Bayes model; (2) train the models; (3) test the models; (4) perform prediction analysis on each of the three models; and (5) compare the performance of the three models using lift charts, classification matrices and profit charts.

Final Project

The final project was in line with the homework assignments. In groups of two or three, students were expected to pick a large data set of their choosing and perform a complete data mining analysis on the data set, produce reports, present a demo, and give a 30 minute presentation of their analysis and findings.

4. DATA SETS

One of the important issues related to teaching a data mining course is to make sure that there are realistic data sets with realistic sizes and enough richness for the purpose of data mining. One data set alone may not have all the characteristics needed for the different data mining tasks. We needed data sets that we can use for the various homework assignments and demos of concepts. We needed a pool of data sets for students to select from for their projects. We needed dirty data sets for students to cleanup, define and build a logical schema, and load into a database. Due to the fact that we needed some data sets with a known "correct" answer, we also found the need to create synthetic data sets where we could control the attribute distributions, dependencies between attributes, and the size of the data set. The following subsections present the various data sets that we used.

Well Known Data Sets

http://www.kdnuggets.com/datasets/ maintains a listing of available data sets. For lectures, homework assignments and hands-on demo(s). We relied heavily on three sources of data, the first source was the UCI data sets that are publicly available. We used the Iris, and Mushrooms data sets. Students used other UCI data sets for their projects. Another source was the AdventureWorks bike shop data from Microsoft, which is in-

cluded as part of their SQL 2005 Server Business Intelligence suite.

Synthetic Data Sets

When introducing new concepts, we found it beneficial to work with datasets that provided known answers – for example, datasets containing very distinctive clusters for cluster analysis or very definite input classifiers for classification. Since most real world datasets generate somewhat fuzzy results, we avoided using real world data when first demonstrating the functionality of a new tool or concept. Instead, we developed an application that generated datasets with predetermined characteristics. The tool allowed us to specify the number of observations desired, number of attributes, the distribution characteristics of attributes, and the relationships between attributes. We also used these datasets for hands-on quizzes and tests where for grading purposes we needed a single correct answer.

Dirty Data Sets

For the purpose of data cleanup and conceptual data modeling, we used chemical toxicity data that is available to us from an on campus environmental science research project. The data is used to extract information for building reference values for mortality, reproduction and growth of species found in areas where contaminants exist. The data was provided in Excel worksheets that containing more than 90,000 records.

5. AVAILABLE RESOURCES

With the advent of the internet and digital media, many faculty around the world have volunteered their lecture notes and video lectures. Companies posted webcasts, books online and tutorials. For the purpose of this paper, we limit our listing to material that was relevant to the course.

Webcasts

As part of their support for SQL Server Business Intelligence tool sets, Microsoft provides a series of webcasts that are available freely for registered users. We found this material helpful for both faculty and students. The Webcasts provide a spectrum of videos ranging from introductory to advanced content (Microsoft Data Mining Webcasts).

Courses Online and tutorials

Faculty, researcher, and company postings of course material for use in a data mining methods course are available. Four sources that we found valuable were:

a) Andrew Moore's tutorials from Auton Lab at Carnegie Mello University (Moore),

b) "Statistical Aspects of Data Mining" a complete course video lecture set from David Mease(2007) - a Google scholar and professor at Stanford University.

c) Microsoft Books online provides step-by-step tutorials, concepts, and whitepapers.

d) videolectures.net provides research oriented content videos. They provide "free and open access to high quality video lectures presented by distinguished scholars". This source is research oriented. It is of value to faculty and graduate students.

6. CONCLUSIONS

We strongly recommend that homework assignments be done in groups of two where students are required (ten minutes block for each group) to demo their homework and present their findings. It emphasizes team work, they learn from each other, and their interpretive skills are nurtured. We do understand that there will be a lot of repetition of content; however, we found that students worked hard to differentiate their work from others who presented ahead of them. As for future offerings of the course, faculty should assign different data sets for different groups where possible, or ask different groups to perform different tasks on the same data set.

We think that there is room for a three-course sequence in data mining methods as follows:

1- An introductory course in Information Theory, Informatics, Information Representation and in-depth XML. The course serves as the entry point to the course sequence. It should have a first course in programming with business statistics as a prerequisite.

2- A first course in Data Mining similar to the course presented in this paper with Database Management and the introductory Information Theory course as prerequisites.

3- An advanced Data Mining Methods course that focuses heavily on the theory of data

mining. In this course, students should build an integrated deployable data mining application using a data mining query language tool kit. This course should include specialized topics such as text mining, support vector machines, etc.

We strongly recommend evaluating "Data Mining Concepts and Techniques" as an alternative textbook (Han 2006). It could serve as a textbook for both the second and third course of the proposed sequence.

One drawback in the course was the students' lack of knowledge of visual interpretation of surface plots. Some background in calculus or a visualization course may help in this area.

The faculty must be very knowledgeable of the underlying technology, frequently we had to provide technical support and troubleshooting of the configuration of the student environment and its connectivity.

The Microsoft Business Intelligence Environment is a self-contained easy to build, configure, and maintain environment. Microsoft provides extensive support in terms of tutorials, books on-line, and video casts. As we now stand, we do not foresee any reason to change the software environment of the course.

7. REFERENCES

Han, J., Kamber M. (2006) Data Mining Concepts and Techniques. Morgan Kaufman Publishers.

Lenox, T. Cuff, Carolyn "Development of a Data Mining Course for Undergraduate Students" ISECON 2002.

Letsche, T. (2007) "Service Learning Outcomes in an Undergraduate Data Mining Course" Proceedings of the 40th Midwest Instruction and Computing Symposium. April 2007.

Lu, Y. Bettine, J. (2002) "Data Mining: An Experimental Undergraduate Course". Journal of Computing Science in Colleges 18, 3 (Feb 2003).

Meesse, D. (2007) http://www.stats202.com/ "Statistics 202: Statistical Aspects of Data Mining".

Microsoft (2007) http://msdn.microsoft.com/en-us/library/ms167488.aspx "SQL Server 2005 Books Online Data Mining Tutorial".

Microsoft (2008) http://www.microsoft.com/sql/technologies/dm/webcasts.mspx "SQL Server 2005 Data Mining Webcasts".

Microsoft (2007) http://www.microsoft.com/sql/technologies/dm/addins.mspx "SQL Server Data Mining Add-Ins for Office 2007".

Moore, A. http://www.autonlab.org/tutorials/ "Statistical Data Mining Tutorials".

North, M. A.; Ahern, T. C.; Fee, S. B. (2007) "The effect of student self-described learning styles within two models of teaching in an introductory data mining course" AESS/IEEE Frontiers in education conference FIE '07. 37th annual.

Pendse, N. (2008) http://www.olapreport.com/ "Consolidation in the BI industry".

Pendse, N. (2007) http://www.olapreport.com/ "Market share Analysis".

Saquer, J. (2007) "A Data Mining Course For Computer Science and Non-Computer Science Students". Journal of Computing Science in Colleges 22, 4 (April 2007).

Tan, B. Steinbach, M. Kumar, V. (2006) Introduction to Data Mining. Pearson Education Inc.

Tang, Z. MacLennan, J. (2005) Data Mining with SQL Server 2005. Wiley Publishing Inc.

UCI (2008) Center for Machine Learning and Intelligent Systems, University of California Irvine Machine Learning Repository http://archive.ics.uci.edu/ml/datasets.html

Witten I., Frank E. (2005) Data Mining Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers.

8. APPENDIX

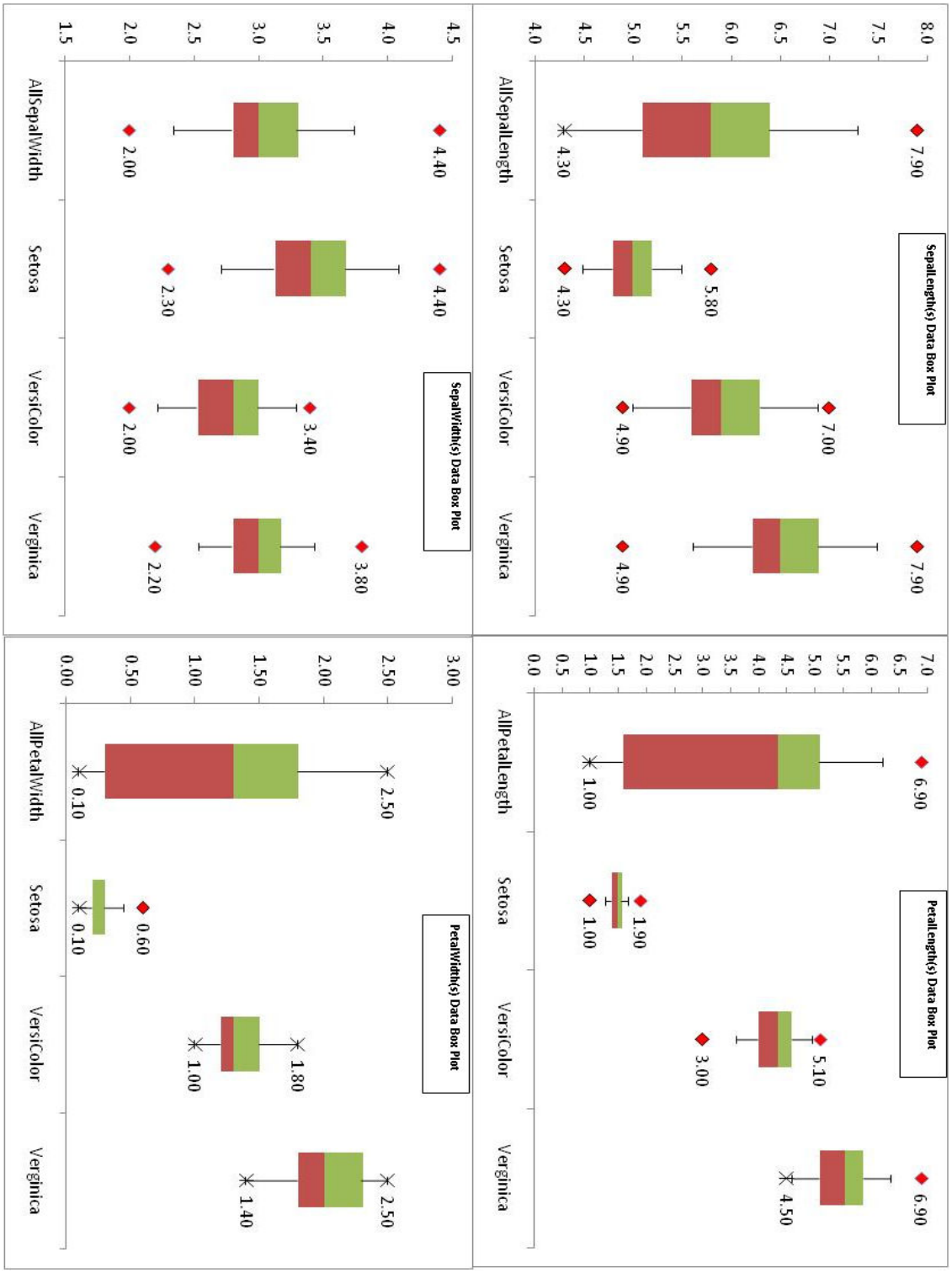

Figure 1 Box Plots of the Various Irises Grouped by Attribute

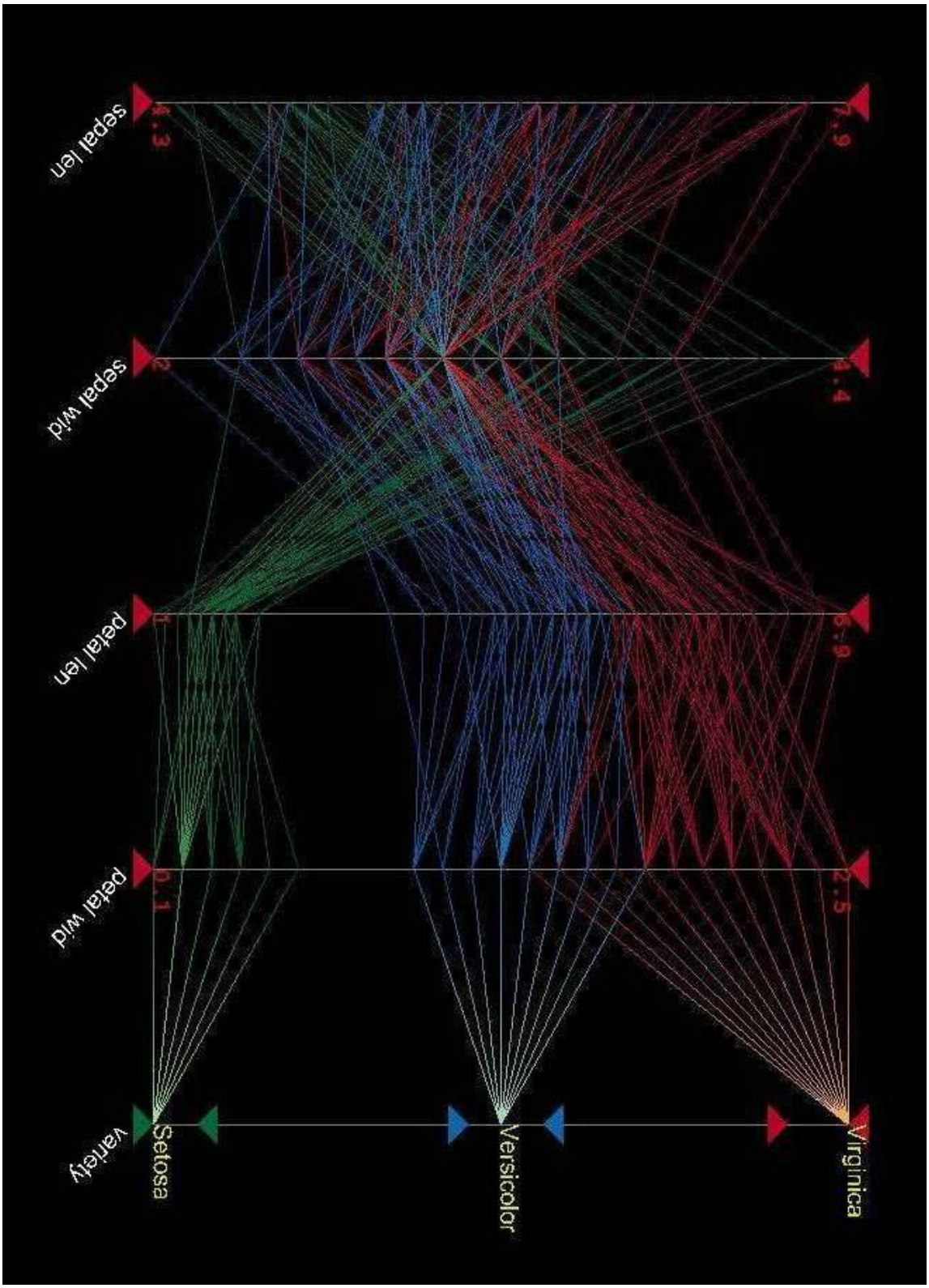

Figure 2 A dynamic Parallel Plot of the Iris Data