Paper 231-26

On the Development of Data Mining Certificate Program at University of Central Florida with the SAS system

Morgan C. Wang, Mori Jamshidian, Ying Zhang, Department of Statistics, University of Central Florida

ABSTRACT

A certificate program in data mining with the aim of training students to master SAS® Enterprise Miner™ software and be ready to joint the data-mining work force upon graduation has established in the Department of Statistics at the University of Central Florida (UCF). This program is strongly supported by SAS Institute and local business community such as Sodexho Marriott and United Space Alliance. SAS training courses covered in this program include SAS Programming I, SAS Program II, SQL Processing with the SAS System, SAS Web Tools: Static and Dynamic Solutions Using SAS/IntrNet®, SAS Web Tools: Advanced Dynamic Solutions, Predictive Modeling Using Logistic Regression, Enterprise Miner: Decision Modeling, Enterprise Miner: Neural Network Modeling, Statistics I: Introduction, and Statistics II: ANOVA and Regression.

INTRODUCTION

The Department of Statistics at University of Central Florida began to offer the graduate level "Data Mining Certificate Program" in fall of 2000. Although this certificate program is still in its infancy, it has been recognized by the leading "data mining software" provider, the SAS Institute. In addition, it has won financial support commitment from local business community such as Sodexho Marriott and United Space Alliance.

Data mining is a relatively new and fast growing industry that has emerged in the 1990s. The META Group, an analyst firm, estimates that the data mining market will grow to $300 million this year and to $800 million by the year 2000. Alex Berson and Stephen J. Smith, two well respected experts in the field of data mining, in their book entitled "Data Warehouse, Data Mining, & OLAP" (page 315) state that

> "The SAS Institute is the largest of the software providers that provide a full range of statistical tools. Founded in 1976, the company has continued to grow from its first statistical products. SAS products are particularly adept at working with large databases for analysis and performing the often required data manipulation and cleansing steps that must be performed before the statistical analysis can begin. They offer perhaps the most comprehensive set of statistical tools to be integrated with OLAP and Data Warehousing and do offer the lower-end JMP software product, which provides a more intuitive and graphical interface to a variety of statistical tools."

To initiate a new certificate program for this relatively new industry, however, requires significant amount of start-up funding from the university administration. In addition, the support from software provider and local business community are also needed. Currently, this program has successfully attracted funding from the following resources:

- Sodexho Marriott: providing up to five fellowship positions each year to support graduate students who are interested in data mining
- United Space Alliance: providing research fund to support faculty to pursuit data mining related research projects
- SAS Institute: providing funding on software and faculty training

In this paper, we address the program, its curriculum, the target audience, faculty resources, and facilities.

PROGRAM DESCRIPTION

The Data Mining Certificate Program provides students the knowledge to use data mining tools, data preparation tools, and data visualization tools needed for data mining with SAS/Enterprise Miner. The completion of this program requires fifteen credit hours, five three-credit-hour courses on data mining methodologies, on statistical analysis, and statistical programming. Two courses, Advanced Computer Processing of Data and Statistical Computing I will build up the SAS programming foundation and prepare students to pass the first two levels of the SAS Certified Professional Exam. Modern computing concepts such as structured programming, database management, object oriented programming, and web based programming will be covered. A third course, Statistical Data Analysis, focuses on statistical methods and data analysis with a significant amount of time devoted to solve modern statistical problems with the SAS system. Final two courses are on data mining methodologies that will cover decision tree, neural network, and logistic regression with state-of-the-arts SAS/Enterprise Miner software. Each course will have a considerable amount of laboratory experiments with data from real world applications such as survey data from UCF Distributed Learning Initiative, maintenance data for solid rocket buster from United Space Alliance, and traffic operation data from Transportation Laboratory at the UCF Civil Engineering Department. All courses will be taught at the main campus of University of Central Florida using state of the art studio style multimedia classrooms.

The descriptions of all the courses offered in this certificate program are summarized below:

- *Advanced Computer Processing Data*
  This course includes data integration, data presentation, information summarization, statistical report writing, and SAS programming fundamentals. Students will learn how to integrate data from different sources such as Microsoft ACCESS databases, web log files, and ASCII files. Producing presentation graphs, statistical tables, and summary statistics with SAS/BASE, SAS/ACCESS®, SAS/GRAPH®, and SAS report writing procedures will be introduced.
- *Statistical Analysis*
  This course includes an overview of statistical methods for analyzing data from experiments and surveys. The emphasis is on detection and modeling of systematic effects of experimental factors, making predictions, and quantifying sources of variation. Specific statistical topics covered include descriptive and graphical methods, t-test and nonparametric test, analysis of categorical data, correlation and regression, multiple regression, analysis of variance and covariance, repeated measure designs, factor analysis, and logistic regression.
- *Statistical Computing I*
  This course uses the SAS system to introduce the concepts of data cleanse, manipulation, and summarization. This concept comprises the data warehouse. The second part of the course shows how to use the SAS system as a tool to perform real

time data analysis and to develop the online information system. SAS/EIS®, SAS/WEBEIS™, SAS/ACCESS®, and SAS/IntraNet will be introduced.

- *Data Mining Methodology I*
  This course is designed to give business decision-makers an overview of the advantages of data mining and to provide analysts with the tools to uncover valuable information through SEMMA (Sample, Explore, Modify, Model and Assess) process. Two modeling techniques including logistic regression and decision tree will be covered. SAS/Enterprise Miner will be introduced.
- *Data Mining Methodology II*
  This course is designed to give business decision-makers in-depth coverage of data mining and provide analysts with additional tools to uncover valuable information through SEMMA (Sample, Explore, Modify, Model and Assess) process. Modeling techniques such as neural network will be covered. Laboratory experiments with SAS/Enterprise Miner and SAS/STAT® will be emphasized.

Upon the completion of the certificate program, students will have gained enough SAS programming experience and statistical knowledge to pass SAS Certified Professional Exams and industrially recognized credentials as an expert of data mining with SAS/Enterprise Miner software.

.

ADMISSION REQUIREMENT

The Graduate Record Exam (GRE or GMAT) is required for all graduate students. The minimum requirements to be considered for admission into the data-mining certificate are

- Grade Point average (GPA) greater than 3.2 for the last 60 credits attempted of credit earned toward the baccalaureate degree
- Graduate Record Exam (GRE) score greater than 1100 on the combined verbal and quantitative sections (GRE score must be less than 5 years old).

Currently, there are twenty-five students from four UCF colleges (College of Arts and Sciences, College of Education, College of Engineering and computer Science, and College of Business) enrolled in this program. In addition, there are university administrators and visiting professors taking courses in this program. The average GRE score for students attending data mining courses is 200 points higher than the minimum requirement.

REQUIRED RESOURCES

Starting this program requires the state of the art software and hardware. The university of Central Florida provided sufficient funding to set up the data mining laboratory and the studio-type classrooms. The SAS Institute provided software and faculty training.

DATA MINING LABORATOTY

This laboratory is located in the Mathematics and Physics building equipped with high-end personal computers, server, and color printer. The lab is open to all students weekdays. If students need to work on their projects during the weekend, they can make arrangement with the lab assistant.

STUDIO STYLE MULTIMEDIA CLASSROOM

Since the instruction is workshop type and includes significant amount of in-class hands on experience type activities, all the courses in the certificate program are held in studio-type meltimedia classrooms. Each classroom has an instructor center that includes one personal computer and one projector. This computer can be used not only for classroom instruction but also for student's activity monitoring. There are twelve personal computers arranged into six stations each with two personal computer and four chairs. Every two students share one personal computer and every four students share a station.

Suppose the instructor wants to show students a feature in SAS Enterprise Miner, the instructor can show students how to use the feature first with his own computer. Students can then use their own computer to explore this new feature just presented by the instructor. If students have any question, the instructor can use the monitor to look at students' problem and provide some help. Since the instructor can monitor all six stations simultaneously, he can provide his help whenever he spots some problems. Teaching in the studio-type classroom can obviously increase the instruction effectiveness.

FACULTY TRAINING

Since the faculties involved in this program do not have enough experience in using Enterprise Miner, the SAS Institute provides free training courses. These training courses include

- Enterprise Miner: Applying Data Mining Techniques
- Predictive Modeling Using Logistic Regression
- Decision Tree modeling
- Neural Network Modeling
- SAS Web Tools: statistic and Dynamic Solutions Using SAS/IntraNet Software
- SAS Web Tools: Advanced Dynamic Solutions Using SAS/IntraNet Software
- SQL Processing with the SAS System

With the training support, the instructors in this program can have in-depth understanding on the newest developed SAS tools on data mining. Consequently, they can provide the up to date information to students in this program.

EXPANSION PLAN

We are pleased that the support received thus far allows us to develop the data mining certificate program. Since the data mining certificate program complement well with the existing Master of Statistical computing program, the department is planning to expand this program to a master degree program on data mining. The master program on data mining will start to offer courses in fall of 2001.

ACKNOLEDGEMENT

The data mining certificate program project is not possible without the help from the following person

- Russell Denslow from Sodexho Marriott
- Randy Raley from Unite Space Alliance
- Jeff Babcock, Frank Lieble, Carolyn Back and DurstSean O'Brien from SAS Institute

REFERENCES

1. SAS Institute Inc., Decision Tree Modeling Course Note, (Cary, NC: author, 2000)
2. SAS Institute Inc., Neural Network Modeling Course Note, (Cary, NC: author, 2000)
3. SAS Institute Inc., Predictive Modeling Using Logistic Regression Course Note, (Cary, NC: author, 2000)
4. SAS Institute Inc., SAS® Programming I: Essentials Course Note, (Cary, NC: author, 2000)
5. SAS Institute Inc., SAS® Programming II: Manipulating Data with the Data Step Course Note, (Cary, NC: author, 2000)
6. SAS Institute Inc., SAS® Web Tools: Running SAS Applications on the Web, (Cary, NC: Author, 1998).
7. SAS Institute Inc., SAS Web Tools: Advanced Dynamic Solutions Using SAS/IntraNet® Software Course Notes, (Cary, NC: Author, 2000)
8. SAS Institute Inc., SAS Web Tools: Static and Dynamic Solutions Using SAS/IntraNet® Software Course Notes,

(Cary, NC: Author, 2000)
9. SAS Institute Inc., SQL Processing with the SAS® System Course Note, (Cary, NC: Author, 2000)
10. SAS Institute Inc., Statistics II: ANOVA and Regression Course Note, (Cary, NC: Author, 2000)

CONTACT INFORMATION

Your comments and questions are valued and encouraged.
Contact author:

Morgan C. Wang
Department of Statistics
University of Central Florida
Orlando, FL 32816-2370
Work Phone: 407-823-2818
Fax: 407-823-3930
Email: cwang@mail.ucf.edu
Web: http://www.pegasus.cc.ucf.edu/~cwang

SAS is a registered trademark or trademark of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.