MACHINE LEARNING ALGORITHMS AND PREDICTIVE MODELS FOR UNDERGRADUATE STUDENT RETENTION AT AN HBCU

A Dissertation

Submitted to the Graduate School

of

BOWIE STATE UNIVERSITY

In partial fulfillment of

the requirements for the

Degree of

DOCTOR OF SCIENCE

Department of Computer Science

By

Ji-Wu Jia

Bowie, MD, 2013

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Machine Learning Algorithms and Predictive Models for Undergraduate Student Retention at an HBC

By

Ji-Wu Jia

DISSERTATION COMMITTEE APPROVAL:

______________________, Chair
Dr. Manohar Mareboyana

____
Dr. Sadanand Srivastava

B. Stone
____
Dr. Daryl Stone

____
Dr. Soo-Yeon Ji

______________________, External Examiner
Dr. Kofi Nyarko

ABSTRACT

Title of Dissertation: MACHINE LEARNING ALGORITHMS AND
PREDICTIVE MODELS FOR UNDERGRADUATE
STUDENT RETENTION AT AN HBCU

Ji-Wu Jia,

Dissertation Chaired By: Dr. Manohar Mareboyana
Department of Computer Science

In this dissertation, I have presented algorithms, which are applied to monitor undergraduate student retention using student data. The study has also made some improvements to the classification algorithms such as Decision tree, Support Vector Machines (SVM), and neural networks that resulted in better prediction accuracies. The experiments revealed that the main factors that influence student retention at the Historically Black Colleges and Universities (HBCU) are the cumulative grade point average and total credit hours taken. The target functions derived from the bare minimum decision tree and SVM algorithms were further revised to create a two-layer neural network and a regression to predict the retention. These new models improved the classification accuracy from 94.03% and 93.64% to 94.42% and 94.29% respectively.

Institutions are increasingly focusing on methods for student retention and graduation. This dissertation examines the data of undergraduate students who continue to be enrolled in an HBCU institution or who have graduated within six years, in order to determine the factors that influence student retention and derive models to predict the retention. To address this issue, the dissertation focuses on generating predictive models of undergraduate student retention using machine learning techniques.

This study applied several machine learning techniques to the student data with a goal to maximize classification accuracy (percentage of instances classified correctly). We further pruned the Weka J48 decision tree, improved the estimated accuracy, and simplified the learning rules for the HBCU undergraduate student retention. After the retention neural networks models were created, the study used learning feedback based performance improvement algorithm to improve the neural networks model's accuracy.

The research presented in the dissertation evaluated the HBCU undergraduate student retention in the six years period and split the six years to seven student academic levels. The study classified and created retention models for each level, and then the dissertation validated the models by seven independent corresponding test datasets. The six-year retention model's out-of-sample error is 6.95%, which is close to the in-sample error 5.58%.

ACKNOWLEDGEMENTS

I would like to first acknowledge the School of Arts and Sciences, the Graduate School, and the Department of Computer Science at Bowie State University. The doctoral program in Applied Science has been a perfect fit for me, as I have learned so much along the way.

Special thanks to Dr. George Acquaah, Dr. Cosmas U. Nwokeafor, Dr. Lethia Jackson, Dr. Hoda El-Sayed, Dr. Manohar Mareboyana, Dr. Sadanand Srivastava, Dr. Darsana Josyula, Dr. Kofi Nyarko, Dr. Daryl Stone, Dr. Soo-Yeon Ji, Dr. Bo Yang, Dr. Joseph Gomes, Dr. Claude Turner, Dr. Joan Langdon, and Dr. Mario Fenyo for their friendly and supportive guidance and assistance throughout this dissertation process. Thanks also to all of my committee members for taking the time to thoroughly review my work and to provide many helpful suggestions and valuable feedback along the way.

Thanks to the Bowie State University Division of Information Technology for their support.

Thanks to Ms. Gayle Fink for her support during my dissertation writing and finally I would like to thank my family for their support.

DEDICATION

This dissertation is dedicated to my wife and my family.

TABLE OF CONTENTS

LIST OF TABLES

Table Title Page

LIST OF FIGURES

LIST OF ACRONYMS

AI	Artificial Intelligence
DM	Data Mining
DT	Decision Tree
EDM	Educational Data Mining
HBCU	Historically Black Colleges and Universities
KDD	Knowledge Discovery in Databases
MDT	Minimum Decision Tree
NN	Neural Networks
NP	Nondeterministic Polynomial Time Complexity
OPAA	Office of Planning, Analysis and Accountability
SVM	Support Vector Machines
VCU	Virginia Commonwealth University
WEKA	Waikato Environment for Knowledge Analysis

CHAPTER I

INTRODUCTION

1. Background of the Problem - Undergraduate Student Retention

Student retention has become an indication of academic performance and enrollment management. One of the biggest challenges that higher education faces is to improve student retention. (Zhang, 2009) Student retention is the Institution's capacity to engage faculty and administrators in a collaborative effort to construct educational settings that engage all students in learning **(**Tinto, 2006). Retention is one of the most widely studied areas in higher education **(**Tinto, 2006). In general, the higher the number of students remaining in the university, the better the academic programs, and the higher the revenue (Zhang, 2009).

Undergraduate students comprise more than 70% of the general university population in the USA. Most studies focus on undergraduate student retention.

Undergraduate student retention is a long-standing problem in higher education and has been the subject of a great deal of research over the past 75 years (Pittman, 2008). Universities with high attrition rates face substantial loss of tuition, fees, and potential alumni contributions, while the students themselves also face negative consequences. Despite the identified consequences of college dropout for universities and students and the concentrated efforts from all educational institutions on improving student retention, attrition rates remain relatively high across the United States. Data from the National

Center for Public Policy and Higher Education reveal that only 73.6% of first-time, full-time freshmen (enrolled in 2002) returned for their second semester. In terms of college completion data from 2005, only 39.5% of undergraduate students enrolled in public institutions completed their degrees within five years. Tinto's widely accepted model of student retention examines factors contributing to a student's decision to continue his or her higher education. The primary focus of this model is a student's academic and social integration into the university (Yu, 2010).

2. Statement of the Problem

Most Historically Black Colleges and Universities (HBCUs) were created at a time when Black students could not attend White institutions. HBCUs were established before 1964, the year the Civil Rights Act outlawed racial segregation, with the intention of serving the African American community (Stone, 2008). African American students are even more likely to drop out of school than their White counterparts.

The problem is that the HBCU needs predictive models for undergraduate student retention to increase the retention rate. This study will investigate, determine, examine, develop, and evaluate the issue being studied. The purpose of this study will be to determine the variables that explain undergraduate student retention at the HBCU. This study is designed to investigate the predictive models for undergraduate student retention at the HBCU.

3. Research Questions and Hypothesis

The dissertation is to answer the following three questions:

1. Can a target function be defined that explains the mapping between student data (academic and social) and six years student retention?

2. Which models can be used to learn different target functions?

3. What is the accuracy of the target function? What methodology can be used to improve the target function's accuracy?

The research hypothesis is as follows:

If we use the algorithms and the modeling target functions, then we can predict the undergraduate student six years retention.

4. Significance of the Study

The HBCU needs predictive models for undergraduate student retention to increase the retention rate. This study will investigate, determine, examine, develop, and evaluate the issue being studied. This study is applying computer science knowledge to the retention problem at the HBCU. The purpose of this study is to explore the effectiveness of machine learning techniques applied to undergraduate student six years retention at the HBCU, and uses those techniques to find patterns of undergraduate students in six years retention at the HBCU. The objective of the study is to reveal algorithms and predictive models of undergraduate student six years retention at the HBCU.

This study is to create predictive models for undergraduate student retention using decision tree and further using Jia Heuristic Algorithm to prune the decision tree to bare minimum decision tree, therefore to obtain the learning rules, and then create neural networks models to predict the retention. This study also created an algorithm for improving the models' accuracy. The problem and the dissertation hypothesis are shown in Figure 1.1.

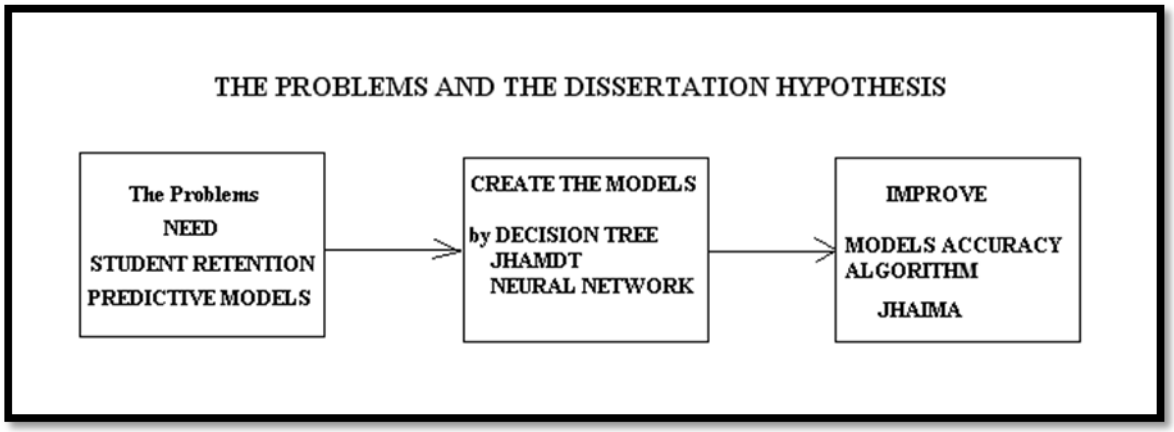

FIGURE 1.1 THE PROBLEMS AND THE HYPOTHESIS

The HBCU can use the study's found algorithms and predictive models to predict undergraduate student retention.

5. Definition of Terms

- Knowledge Discovery in Databases (KDD)

Knowledge discovery in databases (KDD) is the process of finding useful information and patterns in data (Dunham, 2003; Mitchell, 1997; Russell, 2010).

- Data Mining

Data mining is the use of algorithms to extract the information and patterns derived by the KDD process (Dunham, 2003; Mitchell, 1997; Russell, 2010).

Data mining is a form of exploratory data analysis for automatically extracting patterns and relationships from immense quantities of data rather than testing pre-formulated hypotheses. In addition, typical data mining techniques include cross-validation, which is considered a form of re-sampling. The major goal of cross-validation is to avoid over-fitting, which is a common problem when modelers try to account for every structure in one dataset (Yu 2010).

- KDD Process and Data Mining Process

The KDD process consists of the following five steps (Dunham, 2003; Mitchell, 1997; Russell, 2010):

a. Selection: The data needed for the data mining process may be obtained from many different and heterogeneous data sources.
b. Preprocessing: The data to be used by the process may have incorrect or missing data. Erroneous data may be corrected or removed; whereas, missing data must be supplied or predicted.
c. Transformation: Data from different sources must be converted into a common format for processing.
d. Data Mining: This process applies algorithms to the transformed data to generate the desired results.
e. Evaluation: How the data mining results are presented to the users is extremely important.

- Machine Learning

Machine learning is used to adapt to new circumstances and to detect and extrapolate patterns (Russell, 2010).

Machine Learning is the study of computer algorithms that improve automatically through experience. Successful applications range from data mining programs that discover general rules from large databases, to information filtering systems that learn the users' reading preferences, to autonomous vehicles that learn to drive on public highways (Mitchell, 1997).

Knowledge representation stores what it knows or hears (Russell, 2010). Artificial Intelligence (AI) includes many Data Mining techniques such as neural networks and classification. However, AI is more general and involves areas outside traditional data mining. AI application also may not be concerned with scalability as datasets may be small (Dunham, 2003).

Machine learning is the area of AI that examines how to write programs that can learn. In data mining, machine learning is often used for prediction or classification. With machine learning, the computer makes a prediction and then, based on whether it is correct, “learns” through examples, domain knowledge, and feedback (Dunham, 2003).

- Classification

Classification maps data into predefined groups or classes (Dunham, 2003). Dunham defined classification as the following:

Given a database $D = \{t_1, t_2, ..., t_n\}$ of tuples (items, records) and a set of classes $C = \{C_1, ..., C_m\}$, the classification problem is to define a mapping $f : D \rightarrow C$ where each t_i is assigned to one class. A class, C_j, contains precisely those tuples mapped to it; that is, $C_j = \{t_i \mid f(t_i) = C_j, 1 <= i <= n, andt_i \in D\}$.

- Prediction

Many real-world data mining applications can be seen as predicting future data states based on past and current data. Prediction can be viewed as a type of classification. The difference is that prediction is predicting a future state rather than a current state. Prediction applications include flooding, speech recognition, machine learning, and pattern recognition (Dunham, 2003).

- Decision Tree

A decision tree is a predictive modeling technique used in classification, cluster, and prediction tasks. Decision tree uses a “divide and conquer” technique to split the problem search space into subsets (Dunham, 2003). Dunham defined it as the following:

A decision tree is a tree where the root and each internal node are labeled with a question. The arcs emanating from each node represent possible answers to the associated question. Each leaf node represents a prediction of a solution to the problem under consideration.

- Decision Tree Model

Decision tree model is a computational model consisting of three parts (Dunham, 2003):

1. A decision tree as defined above
2. An algorithm to create the tree
3. An algorithm that applies the tree to data and solves the problem under consideration.

These decisions generate rules for the classification of a dataset (Baradwaj, 2011).

- NP (Non-deterministic Polynomial Time)-Hard Problem

Occam’s razor refers to the simplest hypothesis that fits the data (Mitchell, 1997). Minimum set covering is an NP-hard problem (Mitchell, 1997). Finding a minimal decision tree (nodes, leaves, or depth) is an NP-hard optimization problem (Mitchell, 1997).

Class NP means that there are an extremely large number of possibilities for the optimal answer, and we do not have an efficient deterministic algorithm to sift them to find the correct one (McConnell, 2001).

- Entropy Measures Homogeneity of Examples

Entropy characterizes the impurity of an arbitrary collection of examples (Mitchell, 1997; Dunham, 2003). Given a collection *S*, containing positive and negative examples of some target concept, the entropy of *S* relative to this Boolean classification is

$$Entropy(S) = -P_1 \log_2 P_1 - P_2 \log_2 P_2 \qquad (1)$$

Where P_1 is the fraction of positive examples in *S* and P_2 is the fraction of negatives examples in *S*.

Notice that the entropy is 0 if all members of *S* belong to the same class. Note that the entropy is 1 when the collection contains an equal number of positive and negative examples.

For general entropy:

$$Entropy(S) = \sum_j -P_j \log_2 P_j \qquad (2)$$

Where P_j is the fraction of *j* type examples in *S*.

- Information Gain Measures the Expected Reduction in Entropy

A measure of the effectiveness of an attribute in classifying the training data, called information gain, is simply the expected reduction in entropy caused by partitioning the examples according to this attribute (Mitchell, 1997; Dunham, 2003). More precisely, the information gain, Gain(S, A) of an attribute A, relative to a collection of examples S, is defined as