An Artificial Neural Network for Predicting Student Graduation Outcomes

Stamos T. Karamouzis and Andreas Vrettos

Abstract— **Declining student graduation rates is a significant and growing problem in higher education. Students are dropping out from colleges for a variety of reasons and school administrators are scrambling to increase graduation rates. Predicting student graduation is of great value to schools and an enormous potential utility for targeted intervention. Considering the promising behavior of Artificial Neural Networks (ANNs) as classifiers led us into the development, training, and testing of an ANN for predicting student graduation outcomes.**

The network was developed as a three-layered perceptron and was trained using the backpropagation principles. For training and testing various experiments were executed. In these experiments, a sample of 1,407 profiles of students was used. The sample represented students at Waubonsee College and it was divided into two sets. The first set of 1,100 profiles was used for training and the remaining 307 profiles were used for testing. The average predictability rate for the training and test sets were 77% and 68%, respectively.

Index Terms— **Forecasting, Prediction, Retention, Student Success.**

I. Introduction

Declining student graduation rates, a significant and growing problem in the United States, has captured the interest of administrators in both 2-year and 4-year institutes of higher education. According to the department of education, in 2003 the six year graduation rate for students that enrolled in 4-year institutions was 54 percent [1]. The graduation rate in 2-year institutions (community colleges) is even lower. In 2003, the three year graduation rate for community college students was 30 percent down from the 2000 graduation rate of 32 percent [2, 3].

Students are dropping out from colleges for a multitude of reasons such as the cost of education, luck of focus, poor academic background, disengagement in campus activities, etc [4]. School administrators are scrambling to increase graduation rates for a variety of reasons. First of all, it is within the mission of each school to educate students (i.e produce graduates) that become productive members of the community and contribute to the economic prosperity of the nation. Additionally, each school knows that the number of students who drop out translate as loss of revenue for the institution. Another reason that schools strive to increase their graduation rates is that graduation and retention rates are a major component of US News college ranking. Similarly, graduation rates are commonly used as indicators of the school effectiveness by state and federal governments and accrediting agencies. In many cases government funding to the school is directly linked to graduation rates.

From the above it is evident that predicting student graduation is of great value to schools and an enormous potential utility for targeted intervention. Upon the identification of students at-risk support mechanisms such as orientations, advising, mentoring, etc. may be employed to boost student persistence and thus increase graduation rates. The task of prediction can be thought as partitioning students into two classes: "successful" graduates and "unsuccessful" graduates. For a 2-year degree program the term "successful" graduate denotes a student that receives his/her degree within three years. For a 4-year degree program a "successful" graduate is a student that receives his/her degree within six years.

Partitioning of a data set in classes is a very common problem in information processing. We find it in quality control, financial forecasting, laboratory research, targeted marketing, bankruptcy prediction, optical character recognition, etc. Artificial Neural Networks (ANNs) of the feed forward type, normally called multilayer perceptrons (MLPs) have been applied in these areas because they are excellent functional mappers (these problems can be formulated as finding a good input-output map) [5].

Considering the potential utility of predicting student graduation outcomes along with the promising behavior of multilayer perceptrons as classifiers led us into the investigation of ANNs as a tool for predicting student graduation rates. This article presents the development, training, and testing of such a network.

II. Related Work

During the past decade a number of researchers attempted to study student data in order to predict enrollment rates, persistence rates, and/or graduation rates. A number of

Manuscript received June 6, 2007.

S. T. Karamouzis is the Truman & Anita Arnold Chair of Computer & Information Sciences at Texas A&M University – Texarkana, Texarkana, TX 75505 USA (phone: 903-223-3188; fax: 903-223-3189; e-mail: stamos.karamouzis@tamut.edu).

A. Vrettos is an Associate Professor at Waubonsee College, Technology, Mathematics & Physical Sciences Department, Sugar Grove, IL 60554 USA (phone: 630-466-2934; e-mail: avrettos@waubonsee.edu).

statistical methodologies were used including factor analysis [6], multiple regression [7, 8], path analysis [9], and discriminant analysis [10]. Each of these approaches produced various levels of success but in terms of discovering previously undetected patterns are not as effective as the potential of ANNs.

Byers and DesJardins [11] reported the first to attempt to use ANNs in predicting enrollment rates, i.e. which students were likely to be enrolled in a 4-year institution. More relevant to our task at hand is the recent work by Barker et al. [12]. They reported the use of ANNs and Support Vector Machines for classifying successful student graduation rates at a 4-year institution. They used 59 parameters that included demographic, academic, and attitudinal information to describe each student. In a variety of experiments their ANN achieved a best-case prediction rate of 63.4% when tested with the testing set and 67.5% when tested with the training set. Their Support Vector Machines achieved a best-case prediction rate of 63.4% when tested with the testing set and 66.1% when tested with the training set.

Our task at had was even more challenging on two accounts. First of all we are called to reason based on less data for each student (12 parameters versus 59). Additionally, we are called to predict graduation outcomes in a 2-year institution where although the national metric for successful graduation is three years; in reality, studies across the country found that on the average students take four years to complete their 2-year degrees. Our data, network architecture, training methodology, and results achieved are described in the following section.

III. Methodology

A Subjects (data)

The subjects (data) for the present study represented students enrolled at Waubonsee College, a community college located in the State of Illinois, during a 5-year period (Fall of 1997 to Spring of 2002).

Table 1. Parameters in each student profile

	Parameter	Value	Explanation
1	Ethnic Code	Enumerated (1 digit)	One digit code denoting ethnicity
2	Gender	Boolean (1 digit)	One digit code denoting gender
3	Intent Code	Integer (2 digits)	Student’s intention for enrolling in the college
4	Appl Age	Integer (2 digits)	Student’s age at the time of applying to college
5	HS Code	Integer (6 digits)	High school from which the student graduated
6	Disability 14	Boolean (1 digit)	Student needing disability services
7	Disability 17	Boolean (1 digit)	Student needing support services
8	Zip	Integer (5 digits)	Zip code of student’s current address
9	County	Integer (3 digits)	County of student’s current address
10	Grad Age	Integer (2 digits)	Student’s age at high school graduation
11	Srtk Major	Integer (4 digits)	Student’s major at the time of applying to college
12	Success	Boolean (1 digit)	Is the student a “successful” graduate?

At any given time the College maintains a variety of information on each student such as academic and demographic information. Considering only the pre-enrollment information (the student’s information at the time of application to the College) we created a 12-parameter profile for each of the students. Table 1 lists the parameters included in each profile.

In total 1,407 profiles were collected. Randomly, 1,100 of them were used for training the neural network (training set) and the remaining 307 for testing (testing set).

B Architecture

Given the computational capabilities of a multilayer perceptron as a universal pattern classifier a three-layered perceptron was developed. The first layer (input level) comprised of 11 neurons (processing elements) - one for each profile parameter minus the “successful” graduation parameter. The third layer (output level) comprised of 2 neurons - one for denoting “successful” graduation and the other “unsuccessful”. A series of tests was conducted in order to establish the optimal number of processing elements in the second layer (hidden level). The tests indicated that 4 processing elements produced the least error during training.

Each neuron (processing element) is fully connected to every neuron in the following layer. Each neuron accumulates input from the neurons in the prior layer and provides output to neurons in the higher layer. Each neuron's output is calculated based on:

$$y_i(t) = f\,(x_i(t))$$

where $y_i(t)$ is the output of neuron i at time t, $x_i(t)$ is an accumulation of input activity from the neurons at the lower layer. In our network for f we used the hyperbolic tangent function (TanH) as a shifted and scaled sigmoid function. This binds the range of each neuron to between -1 and 1, and often results in increased training efficiency [13], [14].

C Training

Considering that the desired responses of our system are known our perceptron was trained with error correction learning. Denoting $y_i(n)$ the system's response at processing element i at iteration n, and $d_i(n)$ the desired response then for a given input profile an instantaneous error $e_i(n)$ is defined by

ISBN: 978-988-98671-0-2

$$e_i(n) = d_i(n) - y_i(n)$$

Based on the principle of gradient descent learning [15] each weight in the network is adapted by correcting the present value of the weight with a term that is proportional to the present input and error at the weight, i.e.

$$w_{ij}(n+1) = w_{ij}(n) + \eta\, e_i(n)\, x_j(n)$$

where η is a small constant, called the learning rate parameter, $w_{ij}(n)$ is the weight between processing elements i and j at iteration n, and $x_j(n)$ is the input value at processing element j at iteration n.

For updating the weights in our network we used an improvement to the straight gradient descent principle by using a memory term (the past increment to the weight). This is described by the following equation:

$$w_{ij}(n+1) = w_{ij}(n) + \eta\, e_i(n)\, x_j(n) + \alpha\, (\, w_{ij}(n) - w_{ij}(n-1)\,)$$

where α is a small constant, called the momentum.

Training was implemented using batch learning, i.e. first we presented all the patterns that describe the student profiles, then accumulated the weight updates, and at the end we updated the weights with the average weight update. The update of the weights after we present all patters constitutes an epoch. Usually, training takes place over several epochs.

To start the training we used small random values for each weight. For updating the weights in the hidden level we set η to 1.0 and α to 0.7. In the output level η and α were set to 0.1 and 0.7 respectively.

IV. Results

After optimizing the network's structure and training the network within 6000 epochs we tested the network's predictive power on the training data set (i.e on the same 1,100 profiles used to train it). The mean square error achieved was 0.18 and the network was able to correctly classify 148 out of 172 successful graduates (86.04%) and 633 out of 928 unsuccessful graduates (68.21%). Table 2 represents the network's performance when tested with the training data.

Table 2. Performance when tested with training data

Performance	*Successful*	*Unsuccessful*
MSE	0,18	0,18
NMSE	1.35	1.36
MAE	0.32	0.32
Min Abs Error	0.000158226	2.301E-05
Max Abs Error	1.05	1.05
r	0.41	0.41
Percent Correct	86.04	68.21

Because too small errors are an indicator that the network is overtrained (i.e. the network simply memorizes the training set and is unable to generalize the problem) we conducted a series of tests to identify the ideal number of training epochs. The goal was to stop the training at an epoch where the network would produce its maximum generalization power. It turned out that after 6000 training epochs the network produced its best predictive power. When tested with the testing set (307 profiles) it produced a mean square error of 0.22 on both successful and unsuccessful graduates. The network successfully classified 26 out of 37 successful graduates (70.27%) and 179 out of 280 unsuccessful graduates (66.29%). Table 3 represents the network's performance when tested with the testing data.

Table 3. Performance when tested with testing data.

Performance	*Successful*	*Unsuccessful*
MSE	0.22	0.22
NMSE	2.15	2.15
MAE	0.36	0.36
Min Abs Error	0,00	0,00
Max Abs Error	1,05	1,05
r	0,21	0,21
Percent Correct	70.27	66.29

V. Conclusion & Future Work

It appears that predicting graduation rates is a persistent challenge for 2 and 4-year academic institutions. Recognizing that prediction constitutes an excellent first step toward intervention and considering the classification power of ANN's we turned into ANN technology for predicting successful graduation rates. In this article we presented the development, training, and testing of such a network. The network was developed as a three-layered perceptron and was trained using the backpropagation principles. For training and testing various experiments were executed. In these experiments, a sample of 1,407 profiles of community college students was used. The sample was divided into two sets. The first set of 1,100 profiles was used for training and the remaining 307 profiles were used for testing. The average predictability rate for the training and test sets were 72% and 68%, respectively. After comparing our results with results achieved by other researcher we are highly encouraged to continue our work in various directions.

First of all, by expanding the repertoire of parameters included in each student profile we believe that the network will achieve even higher levels of predictability. Additionally, by including more profiles in the training set should refine the network's performance. This is also useful for exploring cross-validation as a method for determining when to stop the network's training in order to avoid overtraining ([16]). Again, the ultimately goal of applying cross-validation would be improved predictive power.

An important follow-up of this work is the identification of the student parameters that are contributory to successful predictability rates. This is of value on two accounts. First of all, it is of value to educational administrators and faculty since if they know the reasons that students do not graduate, within the appropriate time period, then they can directly intervene.

Equally, it is of value in fine tuning the ANN. By eliminating the parameters that are the least contributory to successful predictability we can reduce input noise and thus optimize the network's performance. The identification of the contributory parameters can be done by performing sensitivity analysis [17]. Such analysis can be conducted over the entire training set and will measure the effects of small changes in each of the input parameters as they relate to the output. Input parameters deemed least contributory to the output could be considered superfluous and thus eliminated from the input space

REFERENCES

[1] Associated Press, *U.S. College Drop-out Rates Sparks Concern*, Diverse Online, Sept. 22, 2006.

[2] J. Roueche, L. F. Johnson, S. D. Roueche, *Embracing the tiger: the effectiveness debate and the community college*, Washington: Community College Press, 1997.

[3] American College Testing (Ed.) (2004). [Online] Available: http://www.act.org

[4] J. Merrow, The Undergraduate Experience: Survival of the Fittest, *The New York Times*, April 24, 2005.

[5] R. Lippman, An introduction to computing with neural nets, *IEEE Trans. ASSP Magazine*, 4, 1987, pp. 4-22.

[6] D. Allen, *The hunger factor in student retention: An analysis of motivation.* AIR 1997 Annual Forum Paper. Paper presented at the 37th Annual Forum of the Association for Institutional Research, Orlando, FL., May 1997. (ERIC Document Reproduction Service No. ED 410 865)

[7] P. Tsui, T. Murdock, and L. Mayer, *Trend analysis and enrollment management.* AIR 1997 Annual Forum Paper. Paper presented at the 37th Annual Forum of the Association for Institutional Research, Orlando, FL., May 1997. (ERIC Document Reproduction Service No. ED 410 887)

[8] J. F. Cubeta, N. L. Travers, and B. G. Scheckley, Predicting the academic success of adults from diverse populations. *Journal of College Student Retention, 2*(4), 2001, pp. 295-311.

[9] S. A. Elkins, J. M. Braxton,., and G. W. James, Tinto's separation stage and its influence on first-semester college student persistence. *Research in Higher Education, 41*(2), 2000, pp. 251-268.

[10] W. E. Roweton, *Predicting rural college retention among first-year freshman*, 1994 (ERIC Document Reproduction Service No. ED 370 501).

[11] J. M. Byers, and S. L. DesJardins, Artificial neural networks: A new approach to predicting application behavior. *Research in Higher Education, 43*(2), 2002, pp. 235-258.

[12] K. Barker, T. Trafalis, and T. R. Rhoads, Learning from Student Data, *Proceedings of the 2004 IEEE Symposium on Systems and information Engineering Design*, April 2004, pp. 79-86.

[13] [13] W.S. Stornetta, & B.A. Huberman, An improved three-layer, backpropagation algorithm, *Proceedings of the IEEE First International Conference on Neural Networks*, *2*, 1987, 737-643.

[14] [14] S. I. Gallant, *Neural Network Learning and Expert Systems*, M.I.T. Press, 1993.

[15] D. E. Rumelhart, G. E. Hinton, & R. J. Williams, Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1: Foundations, chapter 8, Cambridge, MA: MIT Press, 1986, pp. 318-362.

[16] M. Stone, Asymptotics for and against cross-validation, *Biometrika, 64*, 1977, 29-35.

[17] A. Saltelli, S. Tarantola, F. Campolongo, & M. Ratto, *Sensitivity Analysis in Practice. A Guide to Assessing Scientific Models*, John Wiley & Sons publishers, 2004.