

Classifiers for Predicting Undergraduate Computer Science Performance

SA Bogle, *Member, IAENG* KM Black

Abstract— This project investigated several learning algorithms to predict first year grades and awards based on input parameters of selected courses which have been cited in the literature as key performance indicators of Computer Science (CS) undergrad performances. The accuracy of predicting first year GPA, was determined by discrete and continuous classifications. Both multiclass and binary classifications were performed for the discrete predications. Naive Bayes, neural network, support vector machine and bagging returned the highest accuracy among discrete classifiers and support vector machine among continuous classifiers. Several learning algorithms were used in the experiments including aggregated methods. Also, the accuracy of the selection test grades in predicting the final class of awards for a bachelor's degree was investigated for a smaller fourth dataset.

Index Terms—: academic performance, classifiers, machine learning, prediction

I. INTRODUCTION

AN aptitude test is administered as a selection test for computer science (CS) student applicants at a Caribbean university. It is intended to extract students with the logic, programming and mathematical aptitude deemed necessary for success in the program. The acceptance ratio is 1 to 6 [1]. Based on concerns over the attrition and transfer rates faced by this program as well as the failure rate in core computing courses, studies have been done to assess the predictive validity of the test and factors affecting performance on the program [2, 3]. This paper focuses on using machine learning approaches to identify the best classifier of predicting performance.

A. Related Work

The attributes used for classification were selected based on the key attribute indicators of computer science performance noted in the literature. For example, [4] asserts that mathematics proficiency is essential to the success of computer science students. Also, studies done by Bergin et

al indicated the other main academic indicators such as Programming [5].

The concern on attrition rates in computer science programs is a longstanding one which has prompted several studies across the decades [6], [7] and [8]. According to [9], the failure rate for first year core computing courses is a cause for concern in several regions. The failure rate for fulltime students in Pre-Calculus was 25% in 2005/2006. In [4], it was asserted that mathematics proficiency at several levels is essential to the success of computer science students. Twenty nine percent of the 2001 applicants failed to complete the full time program in 2005, including six percent of students who were discontinued from the program [3].

II. DATA DESCRIPTION

The data set includes 126 instances of grades for five selected first year subjects namely, Computer Logic & Digital Design, C Programming, Information Technology, Introduction to Programming and Pre-Calculus. These core subjects along with grades from the selection aptitude test were used to predict the first year GPA- the target attribute. Table I outlines how the data was separated to perform the experiments.

The discrete multiclass classification has eleven classifiers of letter grades based on Table II. The accuracy obtained for multiclass classification was significantly lower than the binary classification, but acceptable given the number of classifiers.

The continuous GPA target attribute was discretized into letter grades by converting to percentages and then applying vlookup spreadsheet function to facilitate the multiclass

TABLE I
DESCRIPTION OF DATA SET AND LEARNING ALGORITHMS

Dataset	Attributes	Target Attribute GPA
1	Aptitude Test (6)	Numerical Letter grade (%)
2	First Year Courses(5)	Numerical Letter grade (%)
3	Aptitude Test & First years Courses(11)	Numerical Letter grade (%)
4	Aptitude Test	Class of Honors(4)

Manuscript received March 21, 2018; revised April, 20, 2018.

S. A. Bogle is an alumni of University of Technology, Jamaica and an Assistant Professor with Humboldt State University, Arcata, CA 95519 USA (corresponding author) phone: 707-8263143; fax: 707-826-3140; e-mail: sab30@humboldt.edu).

KM Black is a Senior Lecturer in the School of Computing & Information Technology, University of Technology, Jamaica 237 Old Hope Road Kingston 6, Jamaica (e-mail: kablack@utech.edu.jm).

TABLE II
LETTER GRADE CONVERSIONS OF GPAs

Percentage	Grade
0 -39	U
40 -44	D
45-49	D+
50- 54	C-
55 -60	C
60- 64	C+
65-69	B-
70 -74	B
75 -79	B+
80 -89	A-
90-100	A

classification shown in Table III. Given that the highest accuracy received on multiclass classification was 42% on Naïve Bayes classifier, then a binary classification was done to predict pass or fail. It was also noted the U classifier had 85% accuracy as it has a larger window than the remaining ten classifiers which varies between five and ten.

With Dataset 4, the accuracy of the selection test grades in predicting the final class of awards for a bachelor's degree was investigated for a smaller dataset. The four classes of awards are First Class Honors (F), Upper Second Class Honors (U), Lower Second Class Honors (L), and Pass (P). The challenge with this dataset is that due to attrition, transfers from the program and between different delivery modes of the program, that may lengthen the period of study, there are a high number of students who did not graduate in the prescribed year. Expectation maximization was used to predict target attribute of remaining instances where data is missing. Decision trees and boosting were also applied to the instances with the target attribute specified.

TABLE III
DISCRETE MULTICLASS CLASSIFICATIONS FOR
DATASETS 1, 2 & 3

Data Set ¹	Decision tree	SNB	NB-multinomial	Neural network
Aptitude Test(5)	A:22.2% MAE 0.165 Size 75 Leaves 38	A:22 %	A: 15% MAE 0.1723	A: 22%
Year 1 Courses (6)	A:34% MAE 0.1343 Size 55 Leaves 28	A:42 % MAE 0.128 9	N/A	A: 39%
Aptitude Test & Year 1 Courses (11)	A:35.7% MAE 0.1348 Size: 63 Leaves 32	A:43 % MAE 0.126 5	N/A	A: 41%

Key A- Accuracy, MAE- Mean Absolute Error

TABLE IV
RESULTS OF DISCRETE BINARY CLASSIFICATIONS DATASETS 1, 2 & 3

Method		Aptitude Test (5)	Year 1 Courses (6)	Aptitude test & Year1 Courses
Decision tree	A (%) MAE FP rate	64.3 0.442 0.523	81.7 0.208 0.221	80.1 0.223 0.239
SNB	A (%) MAE FP rate	64.3 0.441 0.419	86.5 0.149 -	86.5 0.1594 0.17
NBmulti-nominal	A (%) MAE	57.9 0.435	n/a	n/a
Neural network	A (%) MAE FP rate	64.3 0.419	88 0.129	88.9 0.11 0.135
AdaBoost Decision tree Stump (default)	A(%) MAE FP rate	68.23 0.439	86.5 1.161 0.166	86.5 0.163 0.166
AdaBoost SNB	A (%) MAE FP rate	57.1% 0.460 -	87.3 0.161 0.144	82.5 0.19 0.21
Bagging RepTree	A (%) MAE FP rate	66.67% 0.428 -	83.3 0.233 0.2	84.1 0.23 0.2
Bagging Naïve bayes	A (%) MAE FP rate	65.8% 0.435 -	86.5 0.153 0.15	85 0.17 0.176
Stacking ZeroR	A (%) MAE	62.7% 0.468	62.7 0.468	62.7 0.4684
SVM	A (%) MAE FP rate	62.7 0.373 -	88.1 0.119 0.174	87.3 0.12 -
Feature Selection	Best First & Greedy Stepwise	Logic	All except IT	Same as Data Set 2

Key A- Accuracy, MAE- Mean Absolute Error

III. RESULTS

A. Discrete Multiclass Classification

All reported results include tenfold cross validation. Table III shows the results obtained when decision trees, simple naïve Bayes (SNB), NB multi-nominal and neural network were applied to the three multiclass datasets: 1-aptitude test, 2- first year courses and 3-aptitude test and year 1 courses combined.

In this experiment, simple Naïve Bayes emerged the best classifier with an accuracy of 43% on the third data set. This is in keeping with Occam's Razor that the simplest can work out to be the best. The low accuracy of 22% on the aptitude test data set is in line with results of statistical studies done by [1] which concluded it was a weak predictor of performance with correlations of approximately 0.2. In [1], higher correlations on the third data set with a maximum of 0.3 were noted. Given the low accuracies of the multiclass classifications, binary classifications were then performed and the results are shown in Table IV. Accuracy, mean

¹ Number of attributes are indicated in parentheses

TABLE V
RESULTS OF CONTINUOUS CLASSIFIERS FOR
DATASETS 1, 2 & 3

Method		Aptitude Test(5)	Year 1 Courses(6)	Year 1 & Aptitude Test
Linear Regression	CC	-0.239	-0.0964	-0.1668
	MAE	0.5696	0.5764	0.5974
	RMSE	0.6997	0.7074	0.7307
Neural Network	CC	0.0227	0.0359	0.0223
	MAE	0.6133	0.6938	0.857
	RMSE	0.7365	0.8876	1.12
Gaussian Processes	CC	-0.0803	-0.0377	-0.1614
	MAE	0.568	0.5561	.5777
	RMSE	0.7053	0.6853	0.7149
SVM for regression	CC	-0.0304	-0.0351	-0.0801
	MAE	0.5677	0.5415	0.58
	RMSE	0.7164	0.7048	0.7383
Regression by disc trees	CC	0.0079	-0.0315	-0.0894
	MAE	0.6275	0.7116	0.7087
	RMSE	0.7652	0.8597	0.8528
Bagging REPtree	CC	0.0034	-0.0946	-0.2366
	MAE	0.5588	0.5626	0.5829
	RMSE	0.6959	0.6914	0.7194
CV Parameter selection	CC	-0.2766	-0.2515	-0.2515
	MAE	0.563	0.5538	0.5538
	RMSE	0.6937	0.6757	0.6757
Stacking	CC	-0.2766	-0.2515	-0.2515
	MAE	0.563	0.5538	0.5538
	RMSE	0.693	0.6757	0.6757
Random Subspace	CC	-0.23	-0.1804	-.02072
	MAE	0.5723	0.5532	0.5633
	RMSE	0.7024	0.6797	0.6873

Key CC – Correlation coefficient, MAE- Mean Absolute Error, RMSE- Root Mean Square Error

absolute error (MAE) and selected false positive (FP) rates are reported.

B. Continuous Classification

Table V shows the results of continuous classifiers such as linear regression and support vector machine as well as bagging and stacking to the three data sets. All results reported include tenfold cross validation for the 126 instances. For the three data sets, linear regression, stacking and CVParameter selection were poor classifiers as they returned the lowest negative correlation coefficients. *Neural networks, Bagging by discretization trees and RepTree returned low positive correlation coefficients and would be considered best classifiers for the first data set.*

Neural network and support vector machines return highest correlations. However, given the higher mean absolute error for neural network, *support vector machine is the best classifier for datasets 2 and 3.*

C. Predicting Classes of Awards

The fourth data set, with Aptitude Test predicting class of honor, was smaller than each of the first three datasets, including only 46 instances and half of the target attribute were missing. Expectation maximization was used to predict missing attributes. The results are shown above in Table VI.

TABLE VI
EXPECTATION MAXIMIZATION CLUSTERED INSTANCES ON
DATASET 4

Class	Distribution
0	12 (27%)
1	3 (7%)
2	7 (16%)
3	10 (22%)
4	13 (29%)

Log likelihood: -21.05325

TABLE VII
ACCURACY & MEAN ABSOLUTE ERROR (MAE) RESULTS ON
DATASET 4: PREDICTING GRADUATING CLASS OF AWARD

Classifier	Accuracy(%) & MAE
SVM	59% 0.303
Decision tree	77% 0.1692
Naïve Bayes	59% 0.2101
Neural Network	59% 0.2147
AdaBoost	73% 0.2383
Decision Stump	59% 0.2898
Bagging	59% 0.2898
REP tree	59% 0.3047
Stacking	59% 0.3047

TABLE VIII
SUMMARY OF BEST DISCRETE CLASSIFIERS FOR
DATASETS 1, 2 & 3

	Multiclass	Binary
Aptitude Test	DT, SNB & NN 22.2%	Boosting DT 68%
Year 1 Courses	SNB 42%	NN & SVM 88%
Aptitude Test and Year 1 courses combined	SNB 43%	NN 89% SVM 87%

Table VII shows the accuracy and mean absolute error of decision trees, neural networks, support vector machine, bagging and stacking on dataset 4.

D. Analysis of results

For the first data set, decision tree, simple Naïve Bayes and neural network all had low accuracies of 22%. The accuracy of just over 40% for the second and third datasets is not wayward in the context of eleven distinct classes for the target attribute. When aggregated methods were applied to the binary classification of the first data set, the accuracy improved by about four percent (See Table III).

Bagging RepTree and Bagging Naïve Bayes also improved the accuracy and were just one or two percent lower than Boosting respectively. Feature selection did not improve the overall accuracy. The features that were extracted from best first and greedy stepwise search were used on decision trees for the third data set and returned 80% accuracy which is

lower than all the other methods used in Table III except stacking which had just over 60% accuracy.

On the continuous data, although neural network had the highest correlation coefficient, its MAE was the highest, and therefore its positive correlation was deemed error prone.

The study by [1] showed a GPA correlation with the entire Aptitude Test of 0.2. This was with a single attribute of the first data set. Overall, the machine learning algorithms applied gave new insights on the data set. In future, best classifiers that emerged from this study can be used on larger data sets of complete cohorts².

The attrition and frequency of change in program modality from full time to part-time and vice versa affected the diversity of the fourth data set. The negative log-likelihood in Table VI is high, and underscores the limit of that dataset. Only one instance had a first class award. These awards are rare, and it would need a significantly larger data set to include more instances with these awards. As shown in Table VII, *decision trees had the highest accuracy of 77% and the lowest mean absolute error*. However, boosting slightly reduced the accuracy. Although the data set is small, *decision trees outperformed all methods in accuracy and reduced error, emerging as the lone best classifier*. Table 8 shows the accuracy of the best discrete classifiers for the first three data sets. For Dataset1, Aptitude Test, the binary classifier had triple the accuracy of the multiclass. For Datasets 2 and 3, the binary classifier had double the accuracy of the multiclass classifier.

IV. CONCLUSION

Aptitude test is a weak classifier of first year performance even when used as a binary classifier. However, when combined with selected year one course results, as done for the third dataset, it boosts the performance. Simple Naïve Bayes, neural networks and decision trees were best classifiers for the first dataset which had the weakest correlation to the target attribute. Simple Naïve Bayes was the best classifier for all three datasets in the Discrete Multiclass classifications. In the binary classifications, neural network and support vector machine emerged best classifiers. Bagging and boosting were aggregated methods that improved the accuracy on the initial experiments. Based on the stacking results, it is a poor classifier for the given data sets, as it reduced the accuracy.

Support vector machine and random subspace are best classifiers for the continuous target attribute GPA. Although neural networks had a positive correlation, the error was significantly higher than other methods.

Decision trees are the best classifier for predicting class of awards, as shown on experiments done on the fourth data set.

V. FUTURE WORK

The aptitude test has since been revised to improve its predictive validity. The first year curriculum has also been revised. A new study could compare results from the revised

aptitude test administered in the last couple of years with the new first year curriculum. More interestingly, gathering data for completed cohorts over an extended time with graduating class of award as the target attribute would be even more meaningful. This would identify new key performance indicators among subjects done over the four year period. A realistic data set for this study would require data from 2012 to 2018 to include as many complete cohorts as possible and mitigate effects of attrition. A longitudinal study of this nature would contain approximately 40 attributes representing grades over a four to five year period. As shown in Table VII, decision trees would be a good candidate classifier for this data set.

REFERENCE

- [1] Bogle, S. & Onyefulu, C. (2009) Assessing the Predictive Validity of an Aptitude Test as a Recruitment Method for Undergraduate Computing. Proceedings of International Conference on Frontiers in Education: Computer Science and Computer Engineering CSREA Press 175-183
- [2] Golding, P. & McNamara, S. (2005). Predicting academic performance in the School of Computing & Information Technology 35th ASEE/IEEE Frontiers in Education Conference S2H-20.
- [3] Golding, P. & Donaldson, O. (2006). Predicting academic performance 36th ASEE/IEEE Frontiers in Education Conference T1D-21
- [4] Tseng CH., Liu JS., Chen YH., Hui L., Jiang YR., Lin JR. (2018) The Requirement Analysis and Initial Design of a Cloud and Crowd Supported Mathematics Learning Environment for Computer Science Students. In: Yen N., Hung J. (eds) Frontier Computing. FC 2016. Lecture Notes in Electrical Engineering, vol 422. Springer, Singapore
- [5] Bergin, S., & Reilly, R. (2005). Programming: Factors that Influence Success. Proceedings of the SIGCSE 2005 Symposium USA:ACM Press.
- [6] Wolfe, J.M., (1971), Perspectives on Testing for Programming Aptitude Proceedings of the 1971 26th annual conference ACM Press
- [7] Campbell, P.F. & McCabe, G.P. (1984). Predicting the Success of Freshmen in a Computer Science Major. Communications of the ACM 27(11) USA:ACM Press.
- [8] Evans, G.E. & Simkin, M.G. (1989). What Best Predicts Computer Proficiency, Communications of the ACM 32(11) ACM Press.
- [9] Moffat, A., Hughes, B., Sundergaard, H., and Gruba, P. (2005). Making Connections: First Year Transition for Computer Science and Software Engineering Students Proceedings of the 7th Australian conference on Computing education 42 Australian Computer Society, Inc.
- [10] Konvalina, J., Wileman, J., & Stephens, (1983). Math Proficiency: A Key to Success for Computer Science Students 26(5) Communications of the ACM USA:ACM Press.

² A cohort refers to a batch of students that start their degree in the same year and complete the program in the specified time frame ie. 4 years fulltime and 5 years part time.