Mining academic data to improve college student retention: An open source perspective

(Research in progress)

Eitel J.M. Lauría, Joshua D. Baron, Mallika Devireddy,

Venniraiselvi Sundararaju, Sandeep M. Jayaprakash

Marist College, Poughkeepsie, NY, USA

ABSTRACT

In this paper we report ongoing research on the Open Academic Analytics Initiative (OAAI), a project aimed at increasing college student retention by performing early detection of academic risk using data mining methods. The paper describes the goals and objectives of the OAAI, and lays out a methodological framework to develop models that can be used to perform inferential queries on student performance using open source course management system data and student academic records. Preliminary results on initial model development using several data mining algorithms for classification are presented.

Categories and Subject Descriptors

J.1 [**Administrative Data Processing**] *Education*; K.3.1 [**Computer Uses in Education**] *Collaborative learning, Computer-assisted instruction (CAI), Computer-managed instruction (CMI), Distance learning*

General Terms

Algorithms, Measurement, Design, Experimentation.

Keywords

Learning Analytics, Open Source, Data Mining, Course Management Systems.

1. INTRODUCTION

Academic or learner analytics has received significant attention within higher education, including being highlighted in the recently released 2011 Horizon Report [4]. This interest can, in part, be traced to the work at Purdue University which has moved the field of academic analytics from the domain of research to practical application through the implementation of Course Signals. Results from initial Course Signal pilots between fall 2007 and fall 2009 have demonstrated significant potential for improving academic achievement [1]. Despite this early success, academic analytics remains an immature field that has yet to be implemented broadly across a range of institutional types, student populations and learning technologies [2]. The Open Academic Analytics Initiative (OAAI), supported by a grant from EDUCAUSE's Next Generation Learning Challenges program, is developing and deploying an open-source ecosystem for academic analytics as means to further research into this emerging field. This paper will focus on two of the five primary objectives of the OAAI: (a) research into the "portability" of predictive models for student performance; and (b) the development and initial deployment of an "open source" model.

LAK'12, 29 April – 2 May 2012, Vancouver, BC, Canada.

To support real-world adoption, OAAI bases its development on open-source technologies already in widespread use at educational institutions, and on established protocols and standards that will enable an even wider variety of existing open-source and proprietary technologies to make use of OAAI code and practices.

The OAAI analyzes student event data generated by Sakai Collaboration and Learning Environment (CLE). The Sakai CLE is an enterprise-level open-source teaching, learning, research, and collaboration software platform initially developed in 2004 by a core group of five institutions (Indiana University, MIT, Stanford University, University of Michigan, and UC Berkeley). Today, Sakai is in use in hundreds of institutions around the world and supported by a vibrant community of developers, designers, educators, and commercial support vendors. Predictive models are developed using the Pentaho Business Intelligence Suite (http://www.pentaho.com/), perhaps the world's most popular open source BI suite, with integrated reporting, dashboard, data mining, and data integration capabilities. Pentaho includes Weka [8], an open source, Java-based sophisticated data mining tool with growing popularity in the data mining community, which is a central piece in the development and testing of predictive models within the OAAI.

An initial set of predictive models for student performance has been developed using Fall 2010 data from Marist's Sakai system, along with student aptitude and demographic data. These models have been deployed using both open-source Weka and IBM's SPSS Modeler to help ensure compatibility between data mining tools. At the conclusion of the OAAI, these predictive models will be released under a Creative Commons/open-source license through the OpenEdPractices.org web site for other institutions to use. The models will be published using the vendor-agnostic XML-based standard Predictive Modeling Markup Language (PMML) which will allow for the importing of models into a range of other analytics tools. Over time this will facilitate an open-source community effort to enhance the predictive models using new datasets from different academic contexts as well as new analytic techniques.

To explore the effectiveness of the Predictive Model for Student Performance, a series of pilots will be run over spring 2012 at three partner institutions, College of the Redwoods (2-year community college), Cerritos College (2-year community college) and Savannah State University (Historically Black College and

University). To support these pilots a Sakai "Student Effort Data" (SED) API has been developed that captures the user activity data Sakai already records to its "event logs" and expose it through a secure standard interface for use by both open-source (Pentaho/Weka) and proprietary external academic analytics tools, such as IBM SPSS Decision Management for Student Performance and SunGard Higher Education Course Signals. Longer-term, Student Information System (SIS) data extraction will be automated and enhanced by leveraging the recently released IMS Global Learning Information Services or LIS standard to facilitate data extraction from Student Information Systems (SIS).

Table 1. Correlations between course grades and CMS data

Undergraduate CMS event frequencies		Course Grade: Marist Fall 2010 N=18968	Course Grade: Campbell (2007) N=27276
Sessions Opened	Correlation	0.147	(no values reported)
	Significance	0.000(**)	
	N	11195	
Content Viewed	Correlation	0.098	0.112
	Significance	0.000(**)	0.000(**)
	N	7651	19205
Discussions Read	Correlation	0.133	0.068
	Significance	0.000(**)	0.000(**)
	N	1552	7667
Discussions Posted	Correlation	0.233	0.061
	Significance	0.000(**)	0.000(**)
	N	1507	7292
Assign. Submitted	Correlation	0.146	0.163
	Significance	0.000(**)	0.000(**)
	N	3245	4309
Asssmts Submitted	Correlation	0.161	0.238
	Significance	0.000(**)	0.000(**)
	N	1423	4085

(**) Significant at the 0.01 level (2-tailed)
Marist data uses ratios over course mean instead of frequencies

One of the initial objectives of the OAAI has been to research into the "portability" of predictive models used in academic analytics to better understand how models developed for one academic context (e.g. large research university), can be effectively deployed in another (e.g. community college). This component of the project is building on the pioneering work of John Campbell whose dissertation research at Purdue University investigated the predictive power of CMS usage data, student aptitude (e.g. SATs and ACT standardized test scores) and student demographic data with regards to student academic success in courses [3]. We have applied similar analytical techniques using Fall 2010 CMS data from Sakai (http://sakaiproject.org), an open-source Course Management System (CMS) started in 2004 and now in use in hundreds of institutions around the world, in production at Marist College to investigate whether similar correlations are found.

Although Marist College and Purdue University differ in obvious ways (e.g., institutional type and size) they do share a number of similarities which are particularly pertinent to this study. These include (2010 data) percentage of students receiving federal Pell Grants (Marist 11%, Purdue 14%), percentage Asian/Black/African American/Hispanic students (Marist 11%, Purdue 11%), and ACT composite 25th/75th percentile (Marist 23/27, Purdue 23/29) [6].

Table 1 shows similarities in correlation values between course grades and CMS frequencies when comparing both institutions. Only comparable metrics between CMS systems (Blackboard in the case of Purdue and Sakai at Marist) have been displayed. As in the case of Purdue, all these metrics are found to be significantly correlated with course grade, with rather low correlation values. Thus, our analysis, as detailed above, provides an initial insight with regards to how "portable" models may be with regards to institutional type and size.

2. METHODOLOGICAL FRAMEWORK

The data mining models considered in our work are based on supervised learning (classification) techniques given that labeled training data is available (data sets used for training purposes carry both input features describing student characteristics and course management system events, as well as student academic performance). The goal is to discriminate between students in good standing and students that are not doing well (a binary classification process)

Our methodological framework consists of six phases, namely Collect data, Rescale/Transform Data, Partition Data, Balance Training Data, Train Models, and Evaluate Models using Test Data. The first four phases deal with preparing the input data used to build (train) and subsequently evaluate (test) models. Trained and tested models can then be used to score incoming data.

Collect Data: Student demographic data and course enrollment data is extracted from the student records system as well as from Sakai (Open source CMS) . Identifying student information is removed during the data extraction process Sakai logs data of individual course events tracked by each of the tools used by an instructor in a given course shell (e.g. Sessions, Content, Discussion Forums, Assignments, Assessments) as well as scores (grade contributions) on gradable events recorded by the Gradebook tool.

Rescale/Transform Data: Data is recoded / processed according to specific needs of the classification model building process. The end product is a data set that collects data of each course taken by each student in a given semester, augmented with student demographic data, CMS event data and partial scores derived from (Gradebook) data. The target (class) variable named Academic Risk establishes a threshold of questionable academic performance (e.g. a cutoff of a C grade or lower defines poor academic performance; a grade above C defines good academic performance).

This stage is also concerned with the removal of outliers, handling of missing data, and addressing the issue of variability among courses in terms of assessment and student activity. The aforementioned variability is dealt with by replacing counts with ratios computed with respect to the average metric for the full course. An aggregated score is derived from partial (Gradebook) scores on gradable events. Once again the purpose is to shave variability across courses and compute a metric that can be used to make early predictions on student academic performance a few weeks into the semester.

Partition Data: input data is randomly divided in two datasets: a training data set, and a test data set. The training data set is used

to build the models. Models are then tested using test data to compute a realistic estimate of the performance of the model(s) on unobserved data. We use a ratio of 70% of the data used for training, and 30% testing, following standard data mining practice.

Balance Training Data: The input data used in the binary classification process is typically unbalanced, as there are usually (many) more students in good academic standing than students at academic risk. In such case where class values are present in highly unequal proportions the number of student-at-academic-risk cases may be too small to render useful information from what distinguishes from good students (the dominant class value). Therefore records of students at academic risk in the training data set are oversampled to level the proportion of classes, and therefore improve the performance of the trained model at detecting such cases. The test data set maintains the original proportions of class values.

Build Predictive Models: We train different models with the training dataset, using different statistical and machine learning processes. We chose three classifiers for comparison purposes: logistic regression, support vector machines, and C4.5 decision trees as they are state of the art robust classification methods that can deal with both categorical and continuous features. Logistic regression is a highly popular parametric classification method, where the target value is a logit function of the linear combination of the predictor features. The *C4.5* [5] *decision tree* algorithm is a non-parametric classifier that learns rules from data. *Support Vector Machines (*SVM) are powerful discriminative models initially proposed by Vapnik [7], that classify data in categories by finding an optimal decision boundary that is as far away from the data in each of the classes as possible.

Evaluate Models: Trained models are evaluated used test data using measures of predictive performance derived from the confusion matrix that yields counts of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Given the unbalanced nature of classes, the overall accuracy (TP+TN)/(TP+TN+FP+FN) is not a good metric for evaluating the classifier, as it is dominated by the student-in-good-standing class (TN+FP). We therefore appeal to two other accuracy metrics: sensitivity (TP/(TP+FN)), which measures the ability of the classifier to detect the class of interest (academic risk); and specificity (TN/(TN+FP)) that measures the number of false alarms raised by the classifier.

3. EXPERIMENTAL SETUP

A data sample corresponding to Fall 2010 undergraduate students was gathered from four different sources: Students biographic data and course related data; Course management (Sakai) event data and Sakai's Gradebook data. Datasets were joined and data was cleaned, recoded, and aggregated to produce an input data file of 3877 records corresponding to courses taken by students. All features considered in the input dataset are listed in Table 2.

FTPT is a flag indicating whether the student is full-time or part-time; ACADEMIC _STANDING identifies probation, regular, or Dean's list students; ENROLLMENT is the course size. RMN_SCORE is the aggregated metric derived from partial (Gradebook) scores described in the previous section. R_SESSIONS and R_CONTENT_READ are Sakai event metrics: R_SESSIONS measures the number of Sakai course sessions opened by a student compared to the course average; R_CONTENT_READ measures the number of content resources viewed by a student compared to the course average.

Table 2. Features (predictors and target) in input dataset

Feature Type	Feature Name
Predictors	GENDER, SAT_VERBAL, SAT_MATH, APTITUDE_SCORE, FTPT, CLASS, CUM_GPA, ENROLLMENT, ACADEMIC _STANDING, RMN_SCORE, R_SESSIONS, R_CONTENT_READ
Target	ACADEMIC_RISK (1 = at risk; 0 student in good standing)

Experiments were conducted using Weka 3.6 and IBM SPSS Modeler 14.2. For each of these tools a flow of execution was developed to perform the experiments. The experiments followed these guidelines:

(i) Out of the input dataset, generate five different random partitions (70% for training, 30% for testing) by varying the random seed

(ii) Balance each training dataset by oversampling records with class ACADEMIC_RISK=1

For each balanced training dataset and each of three classification algorithms (Logistic Regression, C4.5 Decision Tree, SVM), train a predictive model, 5 x 3 = 15 models all in all. For the purpose of this experimental work a radial basis function (RBF) kernel was considered for the SVM algorithm, with a gamma parameter of 0.2; the regularization parameter C was set to 10.

(iii) Using each corresponding test dataset , evaluate each classifiers' performance by measuring their predictive performance (sensitivity, specificity)

(iv) Produce summary measures (mean and standard error)

4. RESULTS

Table 3 displays the assessment of predictive performance of all three classifiers over five different trials. Predictive performance is summarized as a point estimate (mean value) and an error bar.

Both the logistic regression and the SVM algorithms considerably outperform the C4.5 decision tree in terms of their ability to detect students at academic risk: the logistic regression classifier attains a mean sensitivity of 87.67% on the test data set, and the SVM yields 82.60%. This means that these algorithms detect, respectively, 87.67% and 82.60% of the student population at risk.

In terms of specificity, the logistic regression classifier attains a mean value of 89.51% on the test data set, and the SVM yields 90.51%. This means that these algorithms produce, respectively, 10.35% and 9.51% of false positives on the test data. These values are moderately high, especially when compared with the specificity scores of the C4.5 classifier (97.03%, meaning that less that 3% of the test data are false positives)

We performed and assessment of the relative predictive power of the predictors under consideration. This helps to focus the modeling efforts on those predictors that matter most and consider

ignoring those with low predictive power. For logistic regression the RMN_SCORE stands in first place, followed closely by ACADEMIC_STANDING and CUM_GPA; in third place R_SESSIONS and SAT_VERBAL. For the SVM classifier RMN_SCORE occupies the first place, followed by CUM_GPA, ACADEMIC_STANDING, R_SESSIONS and SAT_VERBAL. The Decision tree follows a similar pattern, although the relative difference in predictive performance of the predictors under considerations seems to be minimal. This finding, paired with the low sensitivity values exhibited by the decision tree classifier, demands further analysis. ACADEMIC_STANDING and CUM GPA are typical predictors of academic performance, as descried in the literature. The use of the RMN_SCORE metric as a predictor seems promising if partial grades (final grade contributions of gradable events, such as assignments, or tests) are available at prediction time, but its validity and usefulness require further investigation. CMS events appear to be second tier predictors when compared to the performance metrics described above.

Table 3. Results of the Classification Performance Analysis

Classif Algorithm	Trial	Sensitivity		Specificity	
		Train	Test	Train	Test
Support Vector Machines	1	95.91%	86.15%	90.85%	90.78%
	2	93.94%	77.46%	91.43%	91.35%
	3	92.64%	86.30%	90.47%	88.89%
	4	94.51%	77.78%	91.50%	90.71%
	5	95.83%	85.29%	90.76%	90.81%
	Mean	**94.57%**	**82.60%**	**91.00%**	**90.51%**
	SE	1.37%	4.56%	0.45%	0.94%
C4.5 Decision Trees	1	100.00%	66.15%	99.92%	96.24%
	2	99.39%	61.97%	99.88%	97.67%
	3	100.00%	58.90%	99.88%	96.91%
	4	100.00%	55.56%	99.92%	96.96%
	5	99.40%	54.41%	99.88%	97.37%
	Mean	**99.76%**	**59.40%**	**99.90%**	**97.03%**
	SE	0.33%	4.80%	0.02%	0.54%
Logistic Regression	1	92.98%	86.15%	89.50%	89.17%
	2	89.09%	88.73%	90.06%	90.33%
	3	90.80%	90.41%	89.51%	88.36%
	4	92.07%	83.33%	90.21%	89.33%
	5	91.07%	89.71%	89.55%	91.09%
	Mean	**91.20%**	**87.67%**	**89.77%**	**89.65%**
	SE	1.46%	2.91%	0.34%	1.06%

5. CONCLUSION

This paper reports on the goals and objectives of the Open Academic Analytic, providing a detailed description of the methodology used to develop predictive models in academic analytics using and open source platform. This research derives its motivation from the need of introducing model development approaches that can be used in practical settings to predict academic performance and carry out early detection of students at risk .The methodology presented in this research has been initially applied on real-world data extracted from Marist College and some preliminary results are reported As the project progresses, this analytical framework for academic success will be deployed using data of other institutions and will overtime be enhanced through open-source community collaboration. The goal is to advance our understanding of technology-mediated intervention strategies by investigating the impact that engagement in an online academic support environment has on student success.

We hope that this initiative is imitated by other higher education institutions as a template to facilitate development of predictive models for early detection of academic risk.

6. ACKNOWLEDGEMENTS

This research is supported by EDUCAUSE's Next Generation Learning Challenges, funded through the Bill & Melinda Gates Foundation and The William and Flora Hewlett Foundation. It is also partially supported by funding from the National Science Foundation, award numbers 1125520 and 0963365.

7. REFERENCES

[1] Arnold, Kimberly E. "Signals: Applying Academic Analytics", *EDUCAUSE Quarterly*, Volume 33, Number 1, 2010.

[2] Baepler, P., Murdoch, C.J. (2010, July). Academic Analytics and Data Mining in Higher Education.

[3] Campbell, J. P. (2007). Utilizing Student Data within the Course Management System to Determine Undergraduate Student Academic Success: An Exploratory Study (Doctoral dissertation, Purdue University, 2007). (UMI No. 3287222).

[4] Johnson, L., Smith, R., Willis, H., Levine, A., and Haywood, K., (2011). The 2011 Horizon Report.Austin, Texas: The New Media Consortium.

[5] Quinlan, J.R., C4.5 : programs for machine learning. The Morgan Kaufmann series in machine learning. 1993, San Mateo, Calif.: Morgan Kaufmann Publishers.

[6] U.S. Department of Education, National Center for Education Statistics. Integrated Postsecondary Education Data System, Fall 2010. Retrieved February 15, 2011 from http://nces.ed.gov/collegenavigator.

[7] Vapnik, V.N., The nature of statistical learning theory. 2nd ed. Statistics for engineering and information science. 2000, New York: Springer.

[8] Hall, M., Frank, E., Holmes, G.,Pfahringer, B., Reutemann, P., Witten, I.H. (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, 11(1)

8. AUTHORS' ADDRESSES

Eitel J.M. Lauría (corresponding author), Mallika Devireddy, Venniraiselvi Sundararaju, Sandeep M. Jayaprakash: School of Computer Science and Mathematics, Marist College, Poughkeepsie, NY 12601, USA.

Joshua D. Baron, Senior Academic Technology Officer, Marist College, Poughkeepsie, NY 12601. USA

Email: {Eitel.Lauria, Josh.Baron, Mallika.Devireddy1, Vennirai.Sundararaju1, Sandeep.Jayaprakash1}@marist.edu