Finding Bottlenecks: Predicting Student Attrition with Unsupervised Classifiers

Christopher McKinlay
California State University, Northridge
Northridge, USA
chris.mckinlay@gmail.com

Efunwande Osoba
California State University, Northridge
Northridge, USA
eosoba@gmail.com

Allen Sarkisyan
California State University, Northridge
Northridge, USA
programminglinguist@gmail.com

Seyed Sajjadi
California State University, Northridge
Northridge, USA
seyed.sajjadi.947@my.csun.edu

Carol Shubin
California State University, Northridge
Northridge, USA
carol.shubin@csun.edu

ABSTRACT

With pressure to increase graduation and reduce time to degree in higher education, it is important to identify at-risk students early. Automated early warning systems are therefore highly desirable. In this paper, we use unsupervised clustering techniques to predict the graduation status of declared majors in four departments at California State University Northridge (CSUN), based on a minimal number of lower division courses in each major. In addition, we use the detected clusters to identify hidden bottleneck courses.

Author Keywords

educational datamining; clustering; regression; kmeans

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous; K.3.1. Computer Uses in Education

INTRODUCTION

Policy makers, the public, university administrators, students and their families are concerned about low graduation rates and lengthy times to degree in higher education. At CSUN for example, the median time to degree is six years and the six-year graduation rate is 41% [2]. A related issue is the incidence of major-switching, since re-declaring a major is time-consuming and costly. Approximately 24% of CSUN students re-declare their major, and the plurality of these changes involve departments in the David Nazarian College of Business and Economics (CoBaE) [1]. For this reason we focused our analysis on four departments within the CoBaE.

With an enrollment of over 6000 undergraduate students, CoBaE is one of largest business schools in the nation. CoBaE confers the second most undergraduate degrees at CSUN (behind the College of Social and Behavioral Science), and it has three of the top ten most popular majors (management, finance, and marketing) at CSUN.

In this paper, we trained K-means classifiers on grade data from undergraduate majors in the CoBaE: economics, business law, management, and marketing. Strongly predictive clusters were present in each of the four departments. We found that cluster separation was driven disproportionately by a small number of bottleneck courses. We also found that training classifiers on just the first three classes on the graduation pathway was an effective early detection method.

RELATED WORK

Educational data mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from the educational sphere. The field encompasses various subdomains such as modeling student learning to better optimize performance, to detecting outliers, to developing automated tutoring systems that intelligently adapt lesson plans to the individual learning styles, see [11].

Luan [9] studied clustering aspect of data mining offers comprehensive characteristics analysis of students and offered likelihood estimates for a variety of outcomes, such as transferability, persistence, retention, and success in classes. Al-Radaideh et al. [3] applied classification techniques to determine the main attributes that may affect student performance. Tair and El-Halees [13], used K-means to predict graduate students performance, and overcome the problem of low grades of graduate students. Ayesha, Mustafa, Sattar and Khan [4] have also used K-means clustering to predict student performance in a particular course. Romero, Ventura and Garca [12] described the full process of clustering, classification, statistics and visualization in the context of mining Moodle (e-learning) data. Our current work uses unsuper-

vised clustering methods to address the issue of large scale student behavior.

METHOD

Data Collection and Preprocessing

We obtained academic records containing grade information from declared majors in four departments in the College of Business and Economics at CSUN. The majors we inspected were Economics, Management, Marketing and Business Law. The data spans a fifteen year period between 2000 and 2014containing 13,484 student records in totaland contains only the courses required for each major.

The grade data for each course were encoded with the following normalized GPA scale prior to statistical analysis:

A	A-	B+	B	B-	C+	C	C-	D+	D	D-	F
2.0	1.7	1.3	1.0	0.7	0.3	0.0	-0.3	-0.7	-1.0	-1.3	-2.0

Table 1. Grade encoding scheme

Missing data was encoded as an F in order to facilitate the analysis, since not taking a required course has the same effect as failing it. The datasets were separated by majors, with columns for graduation, number of semesters in the major, number of credits for the major, followed by the course names.

Cluster Analysis

There are a few fundamental issues involved in cluster analysis, notably determining whether discrete clusters are present [8] and choosing the appropriate number of clusters [7] [6]. We applied the K-means algorithm [10] to the grade data and used the Calinski-Harabasz (CH) index [5] to determine the optimal number of clusters on fivefold cross-validated datasets (1).

We then established the predictive power of the clusters by testing them on a classification task. We compared the cluster-based classifiers with logistic regression classifiers in predicting the graduation status of hold-out samples for each department. Receiver Operating Characteristic (ROC) curves were then used to evaluate and compare predictive performance of clustering and logistic regression methods for each department (4). Finally, we performed the same steps to predict graduation status based on the typical first three courses in each major.

RESULTS

The optimal number of clusters was determined to be two in all departments (see figure 1). The Management and Marketing departments showed better between-cluster separation than Economics and Business Law.

We applied the same approach to the first three classes that students would normally take within their first year at school. Table 2 shows the Accuracy, Precision, Recall, F1 Scores and False Omission Rates resulting from the both classifiers when trained on the full feature set and on the first three courses in each major.

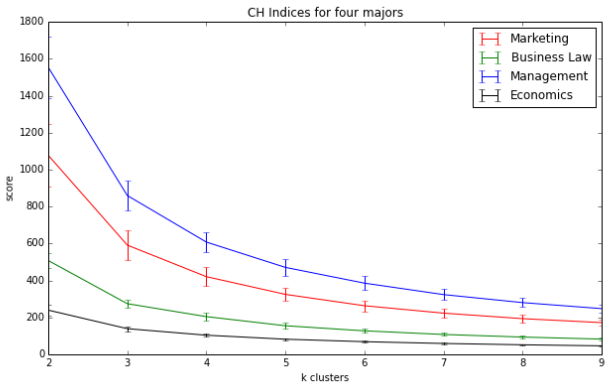

Figure 1. CalinskiHarabasz indices for all four datasets

We expected that a predictive model trained on the full feature set of course grades would be more effective than a model using cluster labels from unsupervised clustering. To test this hypothesis, we compared the performance of a logistic regression classifier trained on the full feature set to the performance of a classifier that used co-membership information from the clusters on a classification task: to predict whether the student had in fact graduated with that major. The cluster-based classifier estimated the probability that a student belonged to a particular category using the fraction of co-clustered samples that also belonged to the category of interest.

In each case we identified strongly predictive clusters. Though outperformed, the cluster-based classifiers compared surprisingly well with the logistic regression models (see figure 4). Heuristically speaking, students in the same cluster tended to drop out at the same times after getting the same grades in the same courses.

DISCUSSION

Cluster analysis can also help to identify common traits among students within each cluster. For each department the second cluster spends on average four semesters enrolled with that major declared (3). However the probability of these students graduating with the major is quite low (see table 2).

	Model	Accuracy	Precision	Recall	F1 Score	False Omission Rate
Management	Logistic Full Set	0.91	0.82	0.91	0.86	0.04
	Logistic 3 Courses	0.7	0.48	0.31	0.38	0.25
	KMeans Full Set	0.78	0.57	1.0	0.73	0.0
	KMeans 3 Courses	0.65	0.46	0.88	0.6	0.08
Economics	Logistic Full Set	0.86	0.81	0.83	0.82	0.1
	Logistic 3 Courses	0.64	0.54	0.45	0.49	0.32
	KMeans Full Set	0.82	0.7	0.94	0.8	0.05
	KMeans 3 Courses	0.45	0.21	0.16	0.18	0.46
Marketing	Logistic Full Set	0.88	0.8	0.89	0.84	0.06
	Logistic 3 Courses	0.68	0.54	0.51	0.52	0.25
	KMeans Full Set	0.78	0.63	0.99	0.77	0.0
	KMeans 3 Courses	0.78	0.63	0.99	0.77	0.0
Business Law	Logistic Full Set	0.79	0.72	0.69	0.71	0.17
	Logistic 3 Courses	0.63	0.5	0.38	0.43	0.32
	KMeans Full Set	0.73	0.58	0.96	0.72	0.04
	KMeans 3 Courses	0.73	0.58	0.96	0.72	0.04

Table 2. Accuracy, Precision, Recall, F1 Scores, and the False Omission Rate

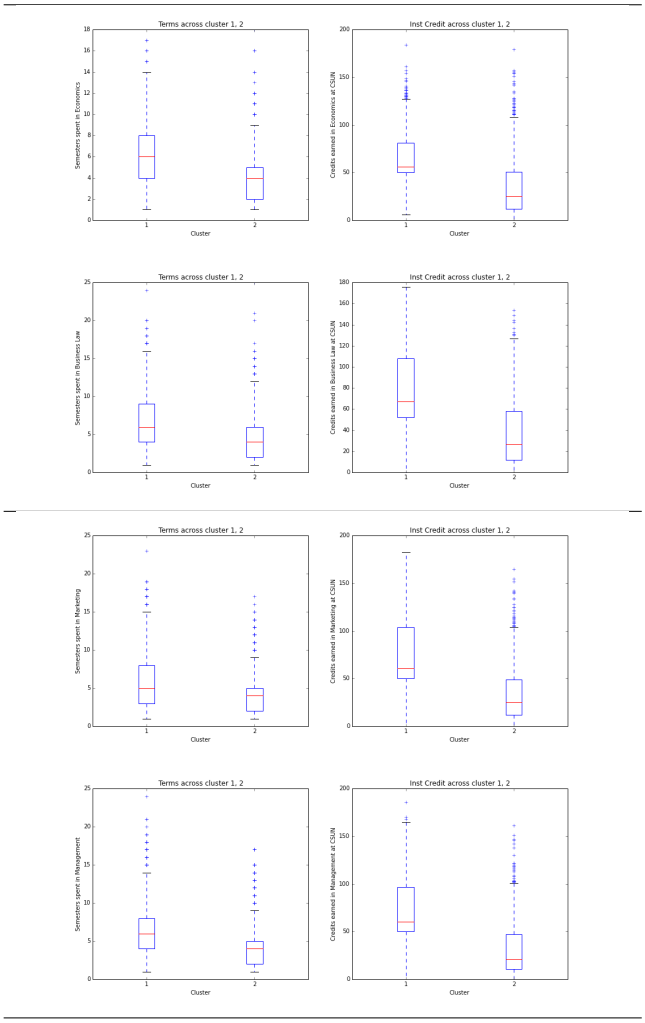

Table 3. Time and units spent in each major

Note that precision in this case is equal to a students probability of graduating given that they are in cluster 1, and the false omission rate is equal to a students probability of graduating given that they are in cluster 2.

Principal component analysis (PCA) plots were used to investigate course contribution to cluster variation for each major. The PCA plots (5) determine the orthonormal directions of maximum variance in each dataset. The student grades are projected onto the first and second principal components, and the red and blue colors mark the clusters as determined by the K-Means algorithm. The management and marketing majors show more spatially separated clusters than those of economics and business law.

A few course vectors are also shown; and here we see that, across majors, the ENGL205 and SOM120 vectors lie parallel to the axis of separation between the two clusters, so these courses do not contribute meaningfully to cluster separation. Conversely, the class vectors (such as ACCT230) that lie perpendicular to the axis of separation contribute the most to cluster separation. Interestingly, there are courses (such as MATH150A) that play this role amongst several different majors. While these courses may not raise flags in any particular department, the fact that they are disproportionately driving student attrition rates should be of concern. We suggest that reforming, or at the very least investigating, these hidden bottleneck courses may be crucial to understanding student attrition at large.

LIMITATIONS AND FUTURE WORK

Students may fail to graduate in CoBaE because they either change majors or discontinue their education at CSUN. Hierarchical clustering methods could provide more detailed information on student outcomes, such as predicting which department a student might change their major to. Collaborative filtering methods could also give departmental recommendations to students considering a change of major. These methods could be used to develop early warning and recommendation systems for automated advisement, which would be especially beneficial to over-taxed advisement systems at comprehensive state universities such as CSUN.

Our results can be further refined by adding student metadata, for example: college year the major was declared, number of transfer credits, number of classes per term, financial aid, student demographics (such as age, gender, ethnicity, zip code) and various measures of student preparedness like SAT scores. With a more detailed feature space, our methods might be able to identify patterns and more well defined clusters.

CONCLUSION

We trained unsupervised classifiers on grade data from four undergraduate majors at CSUN. In each case we found strongly predictive clusters, and found that cluster separation was driven disproportionately by a small number of bottleneck courses. We also found that training classifiers on the first three classes on the graduation pathway was an effective early detection method. We argue that reforming, or at the very least investigating, these bottleneck courses are crucial to understanding student attrition.

ACKNOWLEDGMENTS

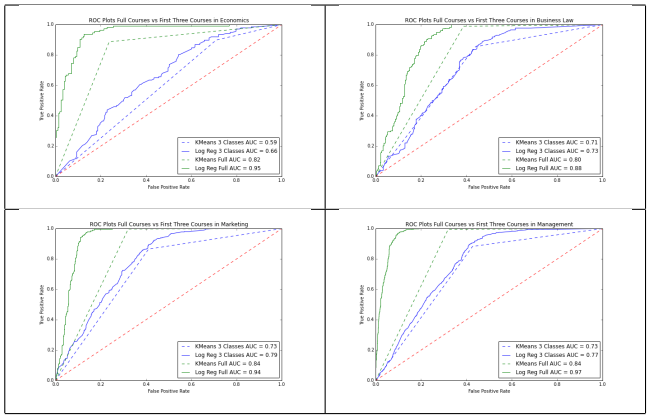

Table 4. ROC plots for both classifiers on the full course set and the first three courses

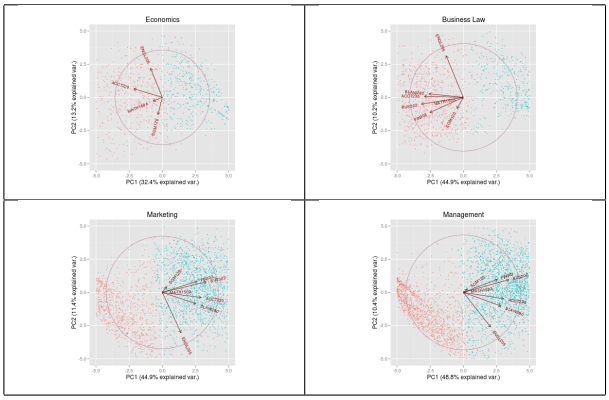

Table 5. PCA plots with course vector labels

We gratefully acknowledge support from CSUNs Office of the Provost and Academic Affairs. We thank Provost Harry Hellenbrand and VIce-Provost Michael Neubauer for their institutional support, and Bettina Huber, CSUN Director of Institutional Research, for making the data available. We also thank Yauheniya (Gina) Lahoda and Dr. Bruce Shapiro for numerous conversations during the Spring 2015 Machine Learning seminar.

REFERENCES

1. CSUN Colleges: Changes and Grades. http://p.rovo.st/blog/?p=47
2. CSUN Office of Institutional Research Report. http://www.csun.edu/~instrsch/retentionrates.html
3. Qasem A Al-Radaideh, Emad M Al-Shawakfa, and Mustafa I Al-Najjar. 2006. Mining student data using decision trees. In *International Arab Conference on Information Technology (ACIT 2006), Yarmouk University, Jordan*.
4. Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, and M Inayat Khan. 2010. Data mining model for higher education system. *Europen Journal of Scientific Research* 43, 1 (2010), 24–29.
5. Tadeusz Caliński and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods* 3, 1 (1974), 1–27.
6. Richard O Duda, Peter E Hart, and David G Stork. 2012. *Pattern classification*. John Wiley & Sons.
7. A.D. Gordon. 1999. *Classification, 2nd Edition*. CRC Press. https://books.google.com/books?id=_w5AJtbfEz4C
8. Dan Knights, Tonya L Ward, Christopher E McKinlay, Hannah Miller, Antonio Gonzalez, Daniel McDonald, and Rob Knight. 2014. Rethinking Enterotypes. *Cell host & microbe* 16, 4 (2014), 433–437.
9. Jing Luan. 2002. Data Mining and Knowledge Management in Higher Education-Potential Applications. (2002).
10. James MacQueen and others. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1. Oakland, CA, USA., 281–297.
11. Cristóbal Romero and Sebastián Ventura. 2010. Educational Data Mining: A Review of the State-of-the-Art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 40, 6 (2010), 601–618.
12. Cristóbal Romero, Sebastián Ventura, and Enrique García. 2008. Data mining in course management systems: Moodle case study and tutorial. *Computers & Education* 51, 1 (2008), 368–384.
13. Mohammed M Abu Tair and Alaa M El-Halees. 2012. Mining educational data to improve students performance: A case study. *International Journal of Information* 2, 2 (2012).