DATA MINING FOR STUDENT RETENTION MANAGEMENT

Shieu-Hong Lin
Department of Mathematics and Computer Science, Biola University
13800 Biola Ave, La Mirada, CA 90639
shieu-hong.lin@biola.edu

ABSTRACT

We conduct a data mining project to generate predictive models for student retention management on campus. Given new records of incoming students, these predictive models can produce short accurate prediction lists identifying students who tend to need the support from the student retention program most. The project is a component in our artificial intelligence class. Students in the class get involved in the entire process of modeling and problem solving using machine learning algorithms. We examine the quality of the predictive models generated by the machine learning algorithms. The results show that some of the machine learning algorithms are able to establish effective predictive models from the existing student retention data.

1. INTRODUCTION

Student retention is a challenging task in higher education [9] and it is reported that about one fourth of students dropped college after their first year [8, 10]. Recent study results show that intervention programs can have significant effects on retention, especially for the first year [7]. To effectively utilize the limited support resources for the intervention programs, it is desirable to identify in advance students who tend to need the support most. In this paper, we describe the experiments and the results from a data mining project in our undergraduate artificial intelligence class to assist the student retention program on campus. The development of machine learning algorithms in recent years has enabled a large number of successful data mining projects in various application domains in science, engineering, and business [4, 11]. In our project, we apply machine learning algorithms to analyze and extract information from existing student data to establish predictive models. The predictive models are then used to identify among new incoming first year students those who are most likely to benefit from the support of the student retention program.

The data mining project serves two purposes. On the one hand, it provides baseline results about the quality of predictive models generated by the machine learning algorithms, which allow the student retention staff to assess the feasibility and utility of incorporating the predictions into the student retention process. On the other hand, it introduces to the artificial intelligence class a real-world application of the machine learning algorithms.

In the remainder of the paper, we describe the details of the project, including the format of the student retention data set, the preprocessing of the data set to protect privacy, the machine learning algorithms applied to the data set, the empirical results on the quality of the predictive models generated by the machine learning algorithms, and the general feedback from the class regarding their experiences in the project.

2. PREPROCESSING THE DATA SET

Students in the artificial intelligence class can access three cleaned-up versions of the student retention data set while participating in the data mining project. They understand the data set is the property of the university and they can not transfer or reveal the data set to others, nor can they use the data set for other purpose. The students also agree to remove the data set from their personal storage after finishing the project.

The raw data set is a collection of 5943 records accumulated over a period of eight years regarding the basic information of first year students and whether they continued to enroll after the first year. Each record in the raw data set keeps track of the values of 52 attributes. In the raw data set, 5009 of the students continued to enroll after their first year while 934 of them dropped out by the end of the first year. We remove attributes involving personally identifiable information such as the name and the student identifier number to protect privacy. We also remove attributes (such as the first-year GPA in college) whose values are not available for new incoming students and are thus useless for student retention prediction. Instead of twelve different attributes regarding various SAT and ACT test scores (many with missing values), we use only a single discretized attribute that summarizes the best test score into one of six discrete levels. Similarly, to make it less likely to be personally identifiable, we only keep discretized attributes that summarize the amounts of financial aid, loan, and scholarship into several discrete levels, and discard the corresponding attributes that record exact amounts. In the end, in the cleaned-up versions of the data set, we only keep twenty two attributes in each record as depicted in Table 1.

Attribute	Description
GENDER	Male or female
ST_RES	State or region from: numeric code
US_CIT	US citizen or not
MAJ_ACAD	Academic major: numeric code
ETH	Ethnic group: numeric code
AGE_c	Age (discretized into 4 levels)
FAFSA	Federal student aid applicant or not
EFCc	Expected family contribution (discretized)
NEED_c	Financial need (discretized)
LOAN_c	Loan received (discretized)
AWD_AMTc	Awarded scholarship (discretized)
CAL_REC	Cal grant receiver or not

Attribute	Description
HOUSE_CoR	Commuter or not
HOUSE_STAT	Living with family or not
HOUSE_ALL	Residence place: numeric code
HS_TYPE	Type of high school attended
GPA_HScnew	High school GPA (discretized)
BES_TESTc	Best test score (discretized)
PER_HSc	Class rank percentile in the high school (discretized)
ACAD_PROB	Under academic probation or not
TYPE_ACPROB	Type of academic probation
RET	Retention: Continue to enroll or not after one year (1 or 0)

Table 1. Attributes used in the cleaned-up versions of the data set

The student retention program needs to focus on students that tend to drop after the first year. However, only less than one sixth of the records in the data set belong to this category. This may mislead some machine learning algorithms to generate models that incline too heavily toward prediction of continued enrollment. To provide a way for balancing such a bias [11], we create two additional versions of the data set by having two or three copies of each of the cases that dropped after a year. Table 2 provides a summary of the raw data set and the cleaned-up versions.

	Description of the contents
Raw data set	5943 records (934 of them dropped after one year, 5009 of them retained), each with 52 attributes.
1x data set	5943 records from the raw data set but only 22 discretized numeric attributes kept.
2x data set	Like the 1x data set, but has 2 copies of each of the original 934 records that dropped after one year.
3x data set	Like the 1x data set, but has 3 copies of each of the original 934 records that dropped after one year.

Table 2. The raw data set and the cleaned-up versions of the data set

3. MACHINE LEARNING FOR FINDING PREDICTIVE MODELS

Weka is open source software that implements a large collection of machine leaning algorithms and is widely used in data mining applications [11]. After learning the conceptual framework of machine learning and the basics of the Weka 3 environment, students participating in the project use Weka to explore the retention data set. They need to conduct experiments on the retention data set to generate predictive models by applying the machine learning algorithms assigned to them. These predictive models (such as those shown in Table 3 below) provide ways to predict whether a new student will continue to enroll or not after one year given the values of the other twenty one attributes.

Simple CART decision tree
derived by the Simple CART algorithm
from the 1x data set

```
GPA_HScnew <= 4
|  NEED_c <= 5
|  |  CAL_REC <= 0       ➔       : 1
|  |  CAL_REC > 0
|  |  |  EFCc_new <= 2  ➔       : 0
|  |  |  EFCc_new > 2   ➔       : 1
|  NEED_c > 5            ➔       : 1
GPA_HScnew > 4           ➔       : 1
```

Alternative decision tree
derived by the alternative decision tree algorithm
(ADT algorithm)
from the 2x data set

```
: 0.493
| (1)GPA_HScnew <= 4          :     -   0.163
| (1)GPA_HScnew > 4           :         0.14
| | (2)BEST_TESTc <= 2        :     -   0.218
| | (2)BEST_TESTc > 2         :         0.135
| (3)NEED_c <= 5              :     -   0.112
| | (4)EFCc_new <= 3          :     -   0.336
| | | (5)NEED_c <= 1          :         0.498
| | | (5)NEED_c >1            :     -   0.384
| | | | (6)EFCc_new <= 2      :     -   0.913
| | | | (6)EFCc_new > 2       :         0.407
| | | | | (7)NEED_c <= 4     :     -   0.927
| | | | | (7)NEED_c > 4      :         0.405
| | (4)EFCc_new >3            :         0.108
| | (10)MAJ_ACAD <= 22        :         0.064
| | (10)MAJ_ACAD > 22         :     -   0.113
| (3)NEED_c > 5               :         0.115
| (8)ST_RES <= 1              :         0.047
| (8)ST_RES > 1               :     -   0.113
| | (9)AGE_c <= 1             :         0.108
| | (9)AGE_c > 1              :     -   0.143
```

J48 graft pruned decision tree
derived by the J48graft algorithm from the 1x data set

```
GPA_HScnew <= 4
|  ST_RES <= 2
|  |  NEED_c <= 5
|  |  |  FAFSA <= -1                                   ➔      : 1
|  |  |  FAFSA > -1
|  |  |  |  EFCc_new <= 2
|  |  |  |  |  NEED_c <= 2                             ➔      : 1
|  |  |  |  |  NEED_c > 2                              ➔      : 0
|  |  |  |  EFCc_new > 2
|  |  |  |  |  EFCc_new <= 4
|  |  |  |  |  |  NEED_c <= 3                          ➔      : 0
|  |  |  |  |  |  NEED_c > 3
|  |  |  |  |  |  |  HOUSE_CoR <= 0
|  |  |  |  |  |  |  |  GENDER <= 1                    ➔      : 1
|  |  |  |  |  |  |  |  GENDER > 1
|  |  |  |  |  |  |  |  |  HS_TYPE4 <= 3               ➔      : 1
|  |  |  |  |  |  |  |  |  HS_TYPE4 > 3
|  |  |  |  |  |  |  |  |  |  LOAN_c <= 1              ➔      : 1
|  |  |  |  |  |  |  |  |  |  LOAN_c > 1
|  |  |  |  |  |  |  |  |  |  |  HOUSE_STAT <= 2       ➔      : 0
|  |  |  |  |  |  |  |  |  |  |  HOUSE_STAT > 2        ➔      : 1
|  |  |  |  |  |  |  HOUSE_CoR > 0                     ➔      : 1
|  |  |  |  |  EFCc_new > 4
|  |  |  |  |  |  NEED_c <= 2
|  |  |  |  |  |  |  EFCc_new <= 6
|  |  |  |  |  |  |  |  NEED_c <= 1                    ➔      : 0
|  |  |  |  |  |  |  |  NEED_c > 1
|  |  |  |  |  |  |  |  |  EFCc_new <= 5               ➔      : 0
|  |  |  |  |  |  |  |  |  EFCc_new > 5                ➔      : 1
|  |  |  |  |  |  |  EFCc_new > 6                      ➔      : 1
|  |  |  |  |  |  NEED_c > 2                           ➔      : 1
|  |  NEED_c > 5                                       ➔      : 1
|  ST_RES > 2
|  |  ETH <= 10                                        ➔      : 1
|  |  ETH > 10
|  |  |  GPA_HScnew <= 3                              ➔      : 0
|  |  |  GPA_HScnew > 3
|  |  |  |  AGE_c <= 1                                 ➔      : 1
|  |  |  |  AGE_c > 1
|  |  |  |  |  MAJ_ACAD <= 24                          ➔      : 0
|  |  |  |  |  MAJ_ACAD > 24                           ➔      : 1
GPA_HScnew > 4                                         ➔      : 1
```

Table 3. Examples of predictive models generated by the machine learning algorithms

Table 3 above shows three decision trees as examples of predictive models learned from the retention data set by three machine learning algorithms: the CART decision tree algorithm [4, 11], the J48 graft decision tree algorithm [4, 11], and the alternative decision tree (ADT) algorithm [4, 11]. For example, consider a new case with a high school GPA of 5 (*GPA_HScnew = 5*), best test score of level 2 (*BEST_TESTc = 2*), financial need of level 6 (*NEED_c = 6*), and is an in-state resident (*ST_RES = 1*). For both the CART decision tree and the J48 graft decision tree [4, 11], we need to start from the root to find a unique path leading to a prediction leaf node. In both trees, we find a unique path of length 1 immediately leading us from the root to a leaf node labeled 1, predicting continued enrollment the next year. On the other hand, for the alternative decision tree (ADT tree), we may have multiple paths from the root to the leaves that are consistent with data and we need to sum up all the numbers appearing on these paths to see whether it is positive or negative [4, 11]. In this particular case, we find three paths leading from the root to leaves. Summing up all the numerical numbers appearing on these paths, we have a positive value 0.493+0.14-0.218+0.115+0.047=0.577, and that leads to the prediction of continued enrollment too. These decision trees also provide interesting insights into hidden patterns in the student retention data set. For example, both the ADT tree and the J48 graft decision tree show that age (attribute *AGE_c*) is a very relevant factor only when the student is not an in-state resident (*ST_RES* > 1) or when the student is an international student (*ST_RES* > 2). Not all predictive models can be visualized conveniently like the decision trees in Table 3. For example, a predictive model generated by the naive Bayes algorithm [4, 11] is simply a collection of statistics derived from the data set while the instance-based nearest neighbor methods [4, 11] essentially match a new case to the entire data set.

In the experiments conducted using Weka 3, we treat all the attributes as numeric attributes and explore the machine learning algorithms applicable to numeric attributes under Weka 3, including (i) fourteen decision tree learning algorithms, (ii) nine decision rule learning algorithms, (iii) four lazy instance-based nearest neighbor algorithms, (iv) seven function-based algorithms for learning neural networks or support vector machines, and (v) five learning algorithms related to the naive Bayes method and Bayesian networks. Typically there are multiple parameters associated with each machine learning algorithm and multiple possible values for each parameter. For each algorithm, we employ mainly the default parameter setting of the learning algorithms in our experiments and have not extensively explored the entire parameter-value space. It is possible that better results can be attained using the same algorithms but with different parameter settings.

4. EVALUATING THE QUALITY OF PREDICTIIVE MODELS

Given the records of new incoming students, we can apply a predictive model to identify students who are likely to drop out if no additional support resources are provided. Staff in the student retention program can more effectively utilize their resources for retention if the predictions are accurate and cover a significant portion of first year students who would drop out if no additional support resources are provided. After a predictive model is established from the student retention data set by a machine learning algorithm, it is then very important to estimate the quality of future predictions generated by the predictive model. In the following, we describe how we conduct cross validation [11] by withholding portions of the student retention data set as test data to evaluate the quality of the predictive model derived by a machine learning algorithm.

Using ten-fold cross validation [11], we first randomly partition the data set into ten subsets, each with 10% of the records in the data set, and then for each subset x we (i) use the algorithm to build a sample predictive model by learning from the other nine subsets combined together and (ii) apply the sample model to predict the retention result for each record in subset x. After the cross validation, we can count to find out four numbers regarding the predictions of all the 5943 records in the student retention data set: (i) n_{dd}, the number of correct predictions among those who dropped after first year, (ii) n_{dc}, the number of wrong predictions among those who dropped after first year, (iii) n_{cc}, the number of correct predictions among those who continued to enroll after first year, (iv) n_{cd}, the number of wrong predictions among those who continued to enroll after first year. We then derive two well known measures, precision and recall [4, 11], from these four numbers to estimate the effectiveness of the predictive model (derived from the entire data set by the algorithm) for identifying new first year students who would drop after a year: (i) precision = $n_{dd}/(n_{dd}+ n_{cd})$ and (ii) recall = $n_{dd}/(n_{dd}+ n_{dc})$. Precision indicates how likely a new case predicted to drop out by the predictive model would actually drop out while recall indicates how likely a would-be drop-out case would be correctly identified by the predictive model.

Table 4 below shows the five machine learning algorithms that produce predictive models with the best precision values in our experiments, together with the corresponding recall values. There is an obvious trade-off between precision and recall when moving from the 1x version of the data set to the 2x version and the 3x version. For these algorithms, the best precision values (ranging from around 68.8% to 84%) are almost all accomplished when learning from the 1x version of the data set, with recall values ranging from 5.1% to 12.3%. Except for the ADT tree, when learning from the 2x version and the 3x version of the data set instead, the precision of the models drops very significantly to the level from 40% to 56.4% while the recall almost all elevates significantly to around the level 27.8% to 88.7%. However, except for the ADT tree and the NB tree [4, 11], the predictive models learned from the 2x version and the 3x version of the data set are huge decision trees involving one thousand nodes or more, unlike the compact decision trees of at most scores of nodes learned from the 1x version of the data set. Given that we only have around six thousand actual records in the data set, decision trees with thousands of nodes seem to overfit the data set and they may not do well as predictive models for predicting new cases [4, 11].

The alternative decision tree (ADT) learning algorithm is the best precision performer we have seen so far, capable of reaching a precision rate of 84% and a recall rate of 12.4% without a sign of overfitting. In other words, given a collection of 1000 new first year students with around 250 would-be drop-out cases embedded in the list (assuming a drop-out rate of 25% according to [8, 10]), the ADT tree algorithm is likely to produce a list of around 37 students and among them about 31 are actual would-be drop-out cases.

	1x data set		2x data set			3x data set		
	Precision	Recall	Precision	Recall	Over-fitting	Precision	Recall	Over-fitting
ADT Tree	83.9%	12.3%	84.0%	12.4%	Unlikely	49.5%	17.6%	Unlikely
NB Tree	77.9%	07.9%	56.4%	08.9%	Unlikely	40.8%	27.8%	Unlikely
CART	73.8%	05.1%	40.4%	49.0%	Likely	44.9%	88.7%	Likely
J48 graft	70.3%	09.6%	44.4%	35.1%	Likely	43.3%	69.8%	Likely
J48	68.8%	09.9%	43.6%	35.2%	Likely	42.7%	69.8%	Likely

Table 4. Precision and recall accomplished by the top predictive models

5. STUDENT FEEDBACK

The students in the artificial intelligence class responded positively regarding their hands-on learning experiences in the project. Compared with small toy data sets not relevant to them, the students indicated from the very beginning that (i) they saw the value of effective predictive models for student retention management and (ii) they felt curious to see whether machine learning algorithms can learn good predictive models from the student retention data set.

Some students indicated that the project helped them to appreciate the importance of machine learning algorithms and they became interested in learning more about the theoretical foundations of machine learning. For students less interested in the theoretical aspect of machine learning, they liked the broad exposure to the entire data mining process. Many of them would like to explore other data mining application domains in the future and felt that the experiences in this project provided a good foundation for their future exploration.

6. CONCLUSIONS

We see some promising empirical results in the preliminary exploration of the data mining project. Machine learning algorithms such as the alternative decision tree (ADT) learning algorithm can learn effective predictive models from the student retention data accumulated from the previous years. The empirical results show that we can produce short but accurate prediction list for the student retention purpose by applying the predictive models to the records of incoming new students.

The benefits of course projects have been well acknowledged in the general context of education [1, 2] and in specific contexts of teaching AI subjects such as stochastic local search [6], case-based reasoning [3], and hidden Markov models [5]. We believe the data mining project described in this paper is another positive example, demonstrating the value of a student project when teaching the theory and practice of machine learning.

REFERENCES

[1] Barron, B. J., Schwartz, D. L., Vye, N. J., Moore, A., Pertrosino, Zech, A., L., Bransford, J., Doing with Understanding: lessons from research on problem- and project-based learning, *Journal of the Learning Sciences*, 7(3-4), 271-311,1998

[2] Blumenfeld, P. C., Soloway, E., Motivating project-based learning: sustaining the doing, supporting the learning, *Educational Psychologist*, 26(3-4), 369-398, 1991.

[3] Bogaerts, S., Leake, D. , Increasing AI project effectiveness with reusable code framework: a case study using IUCBRF, *Proc. 19th International FLAIRS Conference on Artificial Intelligence*, 2-7, 2005.

[4] Han, J., Kamber, M., Pei, J., *Data Mining: Concepts and Techniques*, 4th. Ed. Morgan Kaufmann, 2011.

[5] Lin, S. , An empirical exploration of hidden Markov models: From spelling recognition to speech recognition, *Proc. 19th International FLAIRS Conference on Artificial Intelligence*, 203-208, 2006.

[6] Neller, T. W. , Teaching stochastic local search, *Proc. 19th International FLAIRS Conference on Artificial Intelligence*, 8-14, 2005.

[7] Pan, W., Guo, S., Alikonis, C., Bai, H. , Do intervention programs assist students to succeed in college?: A multilevel longitudinal study, *College Student Journal*, 42, 90-98, 2008.

[8] Tinto, V. , *Leaving College: Rethinking the Causes and Cures of Student Attrition*, University of Chicago Press, 1993.

[9] Tinto, V. , *Research and practice of student retention: What next*, *College Student Retention: Research, Theory, and Practice*, 8(1), 1-20, 2006.

[10] Tinto, V., Russo, P., Kadel, S. , Constructing educational communities in challenging circumstances, *Community College Journal*, 64(1), 26-30,1994.

[11] Witten, I. H., Frank, E., Hall, M. A., *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd. Ed. Morgan Kaufmann, 2011.