Role of Machine Learning in Education: Performance Tracking and Prediction of Students

Dr. B. Subba Reddy[1], S. Shresta[2], S. Sathhvika[2], P. Lakshmi Manasa Shreya[2]

[1]Professor and Head of the Department, [2]UG Scholar, [1,2]Department of Information Technology

[1,2]Malla Reddy Engineering College for Women (A), Maisammaguda, Medchal, Telangana

ABSTRACT

Accurately predicting students' future performance based on their ongoing academic records is crucial for effectively carrying out necessary pedagogical interventions to ensure students' on-time and satisfactory graduation. Although there is a rich literature on predicting student performance when solving problems or studying for courses using data-driven approaches, predicting student performance in completing degrees (e.g. college programs) is much less studied and faces new challenges: (1) Students differ tremendously in terms of backgrounds and selected courses; (2) Courses are not equally informative for making accurate predictions; (3) Students' evolving progress needs to be incorporated into the prediction. In this paper, we develop a novel machine learning method for predicting student performance in degree programs that is able to address these key challenges. The proposed method has two major features. First, a bi-layered structure comprising of multiple base predictors and a cascade of ensemble predictors is developed for making predictions based on students' evolving performance states. Second, a data-driven approach based on latent factor models and probabilistic matrix factorization is proposed to discover course relevance, which is important for constructing efficient base predictors. Through extensive simulations on an undergraduate student dataset collected over three years at UCLA, we show that the proposed method achieves superior performance to benchmark approaches.

Keywords: Education system, Student performance, Machine learning.

1. INTRODUCTION

Making higher education affordable has a significant impact on ensuring the nation's economic prosperity and represents a central focus of the government when making education policies. Yet student loan debt in the United States has blown past the trillion-dollar mark, exceeding Americans' combined credit card and auto loan debts. As the cost in college education (tuitions, fees and living expenses) has skyrocketed over the past few decades, prolonged graduation time has become a crucial contributing factor to the ever-growing student loan debt. In fact, recent studies show that only 50 of the more than 580 public four-year institutions in the United States have on-time graduation rates at or above 50 percent for their full-time students.

To make college more affordable, it is thus crucial to ensure that many more students graduate on time through early interventions on students whose performance will be unlikely to meet the graduation criteria of the degree program on time. A critical step towards effective intervention is to build a system that can continuously keep track of students' academic performance and accurately predict their future performance, such as when they are likely to graduate and their estimated final GPAs, given the current progress. Although predicting student performance has been extensively studied in the literature, it was primarily studied in the contexts of solving problems in Intelligent

Tutoring Systems (ITSs) or completing courses in classroom settings or in Massive Open Online Courses (MOOC) platforms. However, predicting student performance within a degree program (e.g. college program) is significantly different and faces new challenges. First, students can differ tremendously in terms of backgrounds as well as their chosen areas (majors, specializations), resulting in different selected courses as well as course sequences. On the other hand, the same course can be taken by students in different areas.

Since predicting student performance in a particular course relies on the student past performance in other courses, a key challenge for training an effective predictor is how to handle heterogeneous student data due to different areas and interests. In contrast, solving problems in ITSs often follow routine steps which are the same for all students. Similarly, predictions of students' performance in courses are often based on in-course assessments which are designed to be the same for all students. Second, students may take many courses but not all courses are equally informative for predicting students' future performance.

Utilizing the student's past performance in all courses that he/she has completed not only increases complexity but also introduces noise in the prediction, thereby degrading the prediction performance. For instance, while it makes sense to consider a student's grade in the course "Linear Algebra" for predicting his/her grade in the course "Linear Optimization", the student's grade in the course "Chemistry Lab" may have much weaker predictive power.

However, the course correlation is not always as obvious as in this case. Therefore, discovering the underlying correlation among courses is of great importance for making accurate performance predictions. Third, predicting student performance in a degree program is not a one-time task; rather, it requires continuous tracking and updating as the student finishes new courses over time. An important consideration in this regard is that the prediction needs to be made based on not only the most recent snapshot of the student accomplishments but also the evolution of the student progress, which may contain valuable information for making more accurate predictions. However, the complexity can easily explode since even mathematically representing the evolution of student progress itself can be a daunting task. However, treating the past progress equally as the current performance when predicting the future may not be a wise choice either since intuition tells us that old information tends to be outdated. In light of the aforementioned challenges, in this paper, we propose a novel method for predicting student performance in a degree program. We focus on predicting students' GPAs but the general framework can be used for other student performance prediction tasks.

Our main contributions are three-fold.

(1) We develop a novel algorithm for making predictions based on students' progressive performance states. It adopts a bilayered structure comprising a base predictor layer and an ensemble predictor layer. In the base layer, multiple base predictors make local predictions given the snapshot of the student's current performance state in each academic term. In the ensemble layer, an ensemble predictor issues a prediction of the student's future performance by synthesizing the local predictions results as well as the previous-term ensemble prediction. The cascading of ensemble predictor over academic terms enables the incorporation of students' evolving progress into the prediction while keeping the complexity low. We also derive a performance guarantee for our proposed algorithm.

(2) We develop a data-driven course clustering method based on probabilistic matrix factorization, which automatically outputs course clusters based on large, heterogeneous and sparse student course grade data. Base predictors are trained using a variety of state-of-the-art machine learning techniques based on the discovered course clustering results. Specifically, only relevant courses in the same

cluster are used as input to the base predictors. This not only reduces the training complexity but also removes irrelevant information and reduces noise in making the prediction.

(3) We perform extensive simulation studies on an undergraduate student dataset collected over three years across 1169 students at the Mechanical and Aerospace Engineering department at UCLA. The results show that our proposed method is able to significantly outperform benchmark methods while preserving educational interpretability.

2. LITERATURE SURVEY

2.1 Learning factors analysis–a general method for cognitive model evaluation and improvement

AUTHORS: H. Cen, K. Koedinger, and B. Junker

ABSTRACT: A cognitive model is a set of production rules or skills encoded in intelligent tutors to model how students solve problems. It is usually generated by brainstorming and iterative refinement between subject experts, cognitive scientists and programmers. In this paper we propose a semi-automated method for improving a cognitive model called Learning Factors Analysis that combines a statistical model, human expertise and a combinatorial search. We use this method to evaluate an existing cognitive model and to generate and evaluate alternative models. We present improved cognitive models and make suggestions for improving the intelligent tutor based on those models.

2.2 Addressing the assessment challenge with an online system that tutors as it assesses

AUTHORS: M. Feng, N. Heffernan, and K. Koedinger

ABSTRACT: Secondary teachers across the United States are being asked to use formative assessment data (Black &Wiliam, 1998a, 1998b; Roediger & Karpicke, 2006) to inform their classroom instruction. At the same time, critics of US government's No Child Left Behind legislation are calling the bill "No Child Left Untested". Among other things, critics point out that every hour spent assessing students is an hour lost from instruction. But, does it have to be? What if we better integrated assessment into classroom instruction and allowed students to learn during the test? We developed an approach that provides immediate tutoring on practice assessment items that students cannot solve on their own. Our hypothesis is that we can achieve more accurate assessment by not only using data on whether students get test items right or wrong, but by also using data on the effort required for students to solve a test item with instructional assistance. We have integrated assistance and assessment in the Assessment system. The system helps teachers make better use of their time by offering instruction to students while providing a more detailed evaluation of student abilities to the teachers, which is impossible under current approaches. Our approach for assessing student math proficiency is to use data that our system collects through its interactions with students to estimate their performance on an end-of-year high stakes state test. Our results show that we can do a reliably better job predicting student end-of-year exam scores by leveraging the interaction data, and the model based on only the interaction information makes better predictions than the traditional assessment model that uses only information about correctness on the test items.

2.3 Personalized grade prediction: A data mining approach

AUTHORS: Y. Meier, J. Xu, O. Atan, and M. van der Schaar

To increase efficacy in traditional classroom courses as well as in Massive Open Online Courses (MOOCs), automated systems supporting the instructor are needed. One important problem is to automatically detect students that are going to-do poorly in a course early enough to be able to take

remedial actions. This paper proposes an algorithm that predicts the final grade of each student in a class. It issues a prediction for each student individually, when the expected accuracy of the prediction is sufficient. The algorithm learns online what is the optimal prediction and time to issue a prediction based on past history of students' performance in a course. We derive demonstrate the performance of our algorithm on a dataset obtained based on the performance of approximately 700 undergraduate students who have taken an introductory digital signal processing over the past 7 years. Using data obtained from a pilot course, our methodology suggests that it is effective to perform early in-class assessments such as quizzes, which result in timely performance prediction for each student, thereby enabling timely interventions by the instructor (at the student or class level) when necessary.

2.4 Mooc performance prediction via Clickstream data and social learning networks

AUTHORS: C. G. Brinton and M. Chiang

We study student performance prediction in Massive Open Online Courses (MOOCs), where the objective is to predict whether a user will be Correct on First Attempt (CFA) in answering a question. In doing so, we develop novel techniques that leverage behavioral data collected by MOOC platforms. Using video-watching clickstream data from one of our MOOCs, we first extract summary quantities (e.g., fraction played, number of pauses) for each user-video pair, and show how certain intervals/sets of values for these behaviors quantify that a pair is more likely to be CFA or not for the corresponding question. Motivated by these findings, our methods are designed to determine suitable intervals from training data and to use the corresponding success estimates as learning features in prediction algorithms. Tested against a large set of empirical data, we find that our schemes outperform standard algorithms (i.e., without behavioral data) for all datasets and metrics tested. Moreover, the improvement is particularly pronounced when considering the first few course weeks, demonstrating the "early detection" capability of such Clickstream data. We also discuss how CFA prediction can be used to depict graphs of the Social Learning Network (SLN) of students, which can help instructors manage courses more effectively.

2.5 Data mining for adaptive learning in a test-based e-learning system

AUTHORS: Y.-h. Wang, and H.-C. Liao

This study proposes an Adaptive Learning in Teaching English as a Second Language (TESL) for e-learning system (AL-TESL-e-learning system) that considers various student characteristics. This study explores the learning performance of various students using a data mining technique, an artificial neural network (ANN), as the core of AL-TESL-e-learning system. Three different levels of teaching content for vocabulary, grammar, and reading were set for adaptive learning in the AL-TESL-e-learning system. Finally, this study explores the feasibility of the proposed AL-TESL-e-learning system by comparing the results of the regular online course control group with the AL-TESL-e-learning system adaptive learning experiment group. Statistical results show that the experiment group had better learning performance than the control group; that is, the AL-TESL-e-learning system was better than a regular online course in improving student learning performance.

3. PROPOSED SYSTEM

We consider a degree program in which students must complete a set of courses to graduate in T academic terms. Courses have prerequisite dependencies, namely a course can be taken only when certain prerequisite courses have been taken and passed. In general, the prerequisite dependency can be described as a directed acyclic graph (DAG). There can be multiple specialized areas in a program which require different subsets of courses to be completed for students to graduate. We will focus on

the prediction problem for one area in this department. Nevertheless, data from other areas will still be utilized for our prediction tasks. The reason is that data from a single area is often limited while different areas still share many common courses.

3.1 ADVANTAGES OF PROPOSED SYSTEM

- It is important for constructing efficient base predictors.
- System that can continuously keep track of students' academic performance and accurately predict their future performance

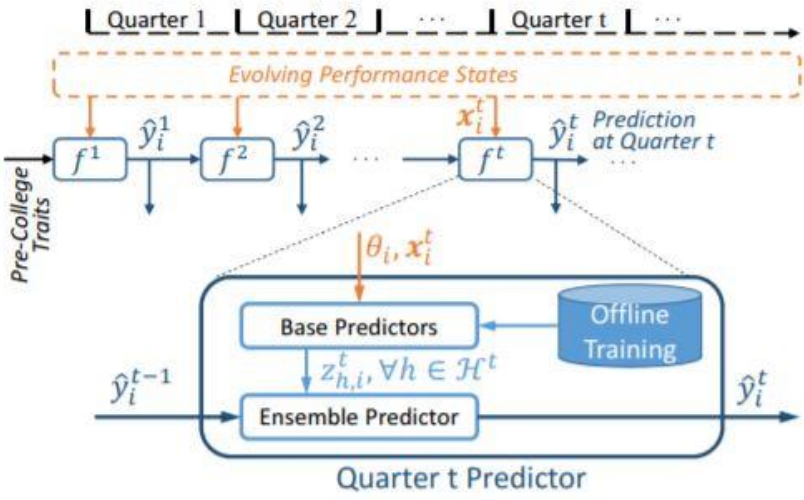


Fig. 2: Proposed student tracking system.

3.2 SYSTEM MODULE

In this work, Machine Learning algorithms are used for suggesting concept to predict future courses performances of students by using students previous terms result data as feature vectors. Every year in all universities only 50 % students completing graduation courses successfully and remaining students are failed to complete course so by using this paper machine learning algorithms college peoples can predict future performance of students by giving his past performance GPA as input to the machine learning algorithms.

To implements this work, we gave solutions to 3 problems:

1. All students will differ in selected courses as different students may take different courses then how we can build machine learning model to predict performance of particular student in particular course. To solve this issue author is using Clustering concept using Matrix Factorization. In matrix factorization we will form cluster of related courses and then in matrix we will put value 1 (or marks obtained by students in term) if user selected course otherwise we will put value 0. So, by forming this cluster we will have matrix feature vector which contains similar courses student's data. This matrix features will be passed to machine learning to train model and to predict future performances. A single course consists of many subjects and all those subjects' marks will be used to calculate GPA in all terms. Past course GPA and marks will be input to machine learning algorithms to predict future course performance.
2. Courses are not equally informative for making accurate predictions and by forming feature vector we will get prediction for selected course only.

3. All existing algorithms were concentrating on past data to predict future performance but in this paper we will used past data as well as ongoing course performance data to predict future course GPA. Students' evolving progress needs to be incorporated into the prediction.

   To implement this work, we have used base classifiers such as Random Forest, SVM, Logistic Regression or KNN. The prediction results of base classifier will be pass to ensemble classifier to predict better results for ongoing courses and future courses.

3.2.1 Implementation module

- Upload UCLA Students Dataset
- Pre-process Dataset
- Feature Extraction
- Model Generation
- Matrix Factorization
- Run SVM Algorithm
- Run Random Forest Algorithm
- Run Logistic Regression Algorithm
- Propose Ensemble based Progressive Prediction (EPP) Algorithm
- Predict Performance
- Mean Square Error Graph

4. RESULTS

4.1 Dataset description

The dataset is considered from UCLA University and this dataset saved inside dataset folder. Below is some example of dataset records.

Math33A,Math33B,Math31B,MATH32A,MAE105A,CHEM20B,MAE103,MATSCI104,MAE105D, MAE94,PHYS1A,PHYS1B,GPA

-0.13,-1.58,0,1.93,0,0,371.08,435.04,717.61,819.58,1.34,1.21,0

0,0,-5.16,1.01,-4.04,-0.49,336.5,481.63,672.11,850.33,4.59,1.68,0

0.84,2.85,1.85,5.32,1.47,1.8,379.72,411.92,586.72,717.06,5.87,1.95,1

0,0,-5.23,0,-4.21,-0.81,403.64,621.58,780.255,768.79,3.26,1.44,1

Above dataset is from UCLA mechanical students and in above dataset all bold names are the subject names of mechanical course and all decimal values are the marks obtained by students in those subjects. If student not taken course then marks will be 0 and in last column GPA is defines as 0 and 1. 0 means low GPA score and 1 means high GPA score, so using above dataset we will build base learning classifier and then pass result to propose ensemble algorithm called Ensemble-based Progressive Prediction (EPP). EPP algorithm will predict GPA value as LOW or HIGH for given test data or new student ongoing performance data. Below is the test data using for new students and in test data there will be no GPA value and EPP algorithm will predict it.

Math33A,Math33B,Math31B,MATH32A,MAE105A,CHEM20B,MAE103,MATSCI104,MAE105D, MAE94,PHYS1A,PHYS1B0

0.27,1.45,1.03,1.69,1.04,0.08,399.67,441.83,704.345,0,0,0

1.31,0.38,-1.78,1.13,-0.64,-1.23,0,477.21,0,630.17,3.78,0

0.51,1.45,-0.24,2,0.44,-0.49,330.89,469.75,688.65,1103.04,3.26,3.13

1.06,1.45,1.03,1.2,2.77,0,0,490.38,592.565,0,2.85,0

In above test dataset only subjects names and its marks are there but no GPA is available and to predict GPA we will apply EPP algorithm. In above dataset if student not taken subject or marks yet to expect then we will put 0.

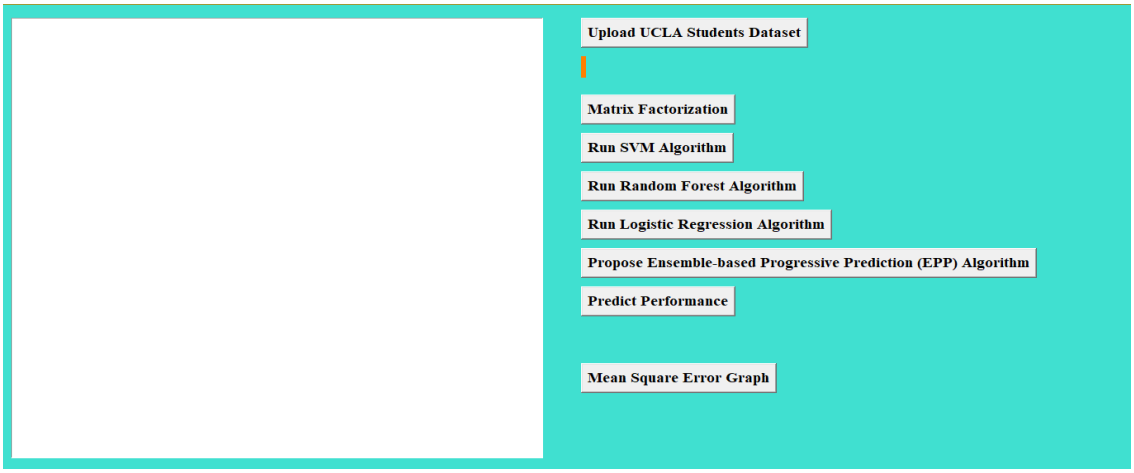
In code we are using below statements to pass base classifier result to ensemble classifier

base = RandomForestClassifier() //creating base classifier object

epp = BaggingClassifier(base_estimator=base)// passing base classifier to ensemble

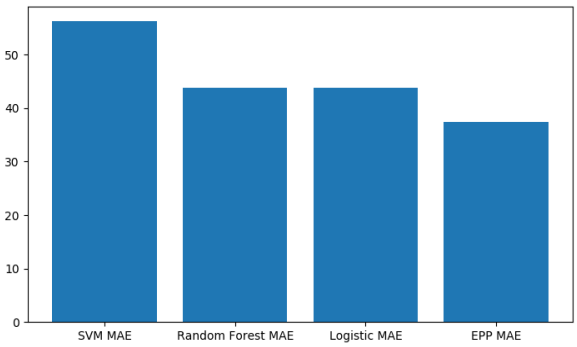
epp.fit(X_train1, y_train) //now training ensemble with past data as training

prediction_data = prediction(X_test1, epp) //now calling prediction function to predict future course GPA by passing test data as student ongoing course result. Prediction will be done using propose EPP algorithm.



In above screen click on ‘Upload UCLA Students Dataset’ button to upload dataset

In below screen in square brackets are the students marks of ongoing subjects and this marks are converted to matrix factorization and then applied on EPP train model to predict GPA as LOW or HIGH. In above screen after each test record I am displaying predicted result value. Now click on “Mean Square Error Graph’ button to get below graph



In above graph x-axis represents algorithm name and y-axis represents MSE (mean square error). From above graph we can see propose algorithm got less MSE error and has high accuracy compare to other algorithms. From above graph we can conclude that propose EPP is better in prediction compare to other algorithms

5. CONCLUSION

In this project, we proposed a novel method for predicting students' future performance in degree programs given their current and past performance. A latent factor model-based course clustering method was developed to discover relevant courses for constructing base predictors. Ensemble-based progressive prediction architecture was developed to incorporate students' evolving performance into the prediction. These data-driven methods can be used in conjunction with other pedagogical methods for evaluating students' performance and provide valuable information for academic advisors to recommend subsequent courses to students and carry out pedagogical intervention measures if necessary. Additionally, this work will also impact curriculum design in degree programs and education policy design in general. Future work includes extending the performance prediction to elective courses and using the prediction results to recommend courses to students.

REFERENCES

[1] The Whitehouse, "Making college affordable,"https://www.whitehouse.gov/issues/education/higher-education/ making-college-affordable, 2016.

[2] Complete College America, "Four-year myth: Making college more affordable," http://completecollege.org/wp-content/uploads/2014/ 11/4-Year-Myth.pdf, 2014.

[3] H. Cen, K. Koedinger, and B. Junker, "Learning factors analysis–a general method for cognitive model evaluation and improvement," in International Conference on Intelligent Tutoring Systems. Springer, 2006, pp. 164–175.

[4] M. Feng, N. Heffernan, and K. Koedinger, "Addressing the assessment challenge with an online system that tutors as it assesses," User Modeling and User-Adapted Interaction, vol. 19, no. 3, pp. 243–266, 2009.

[5] H.-F. Yu, H.-Y. Lo, H.-P. Hsieh, J.-K. Lou, T. G. McKenzie, J.-W. Chou, P.-H. Chung, C.-H. Ho, C.-F. Chang, Y.-H. Wei et al., "Feature engineering and classifier ensemble for kdd cup 2010," in Proceedings of the KDD Cup 2010 Workshop, 2010, pp. 1–16.

[6] Z. A. Pardos and N. T. Heffernan, "Using hmms and bagged decision trees to leverage rich features of user and skill from an intelligent tutoring system dataset," Journal of Machine Learning Research W & CP, 2010.

[7] Y. Meier, J. Xu, O. Atan, and M. van der Schaar, "Personalized grade prediction: A data mining approach," in Data Mining (ICDM), 2015 IEEE International Conference on. IEEE, 2015, pp. 907–912.

[8] C. G. Brinton and M. Chiang, "Mooc performance prediction via clickstream data and social learning networks," in 2015 IEEE Conference on Computer Communications (INFOCOM). IEEE, 2015, pp. 2299– 2307.

[9] KDD Cup, "Educational data minding challenge," https://pslcdatashop. web.cmu.edu/KDDCup/, 2010.

[10] Y. Jiang, R. S. Baker, L. Paquette, M. San Pedro, and N. T. Heffernan, "Learning, moment-by-moment and over the long term," in International Conference on Artificial Intelligence in Education. Springer, 2015, pp. 654–657.

[11] C. Marquez-Vera, C. Romero, and S. Ventura, "Predicting school failure using data mining," in Educational Data Mining 2011, 2010.

[12] Y.-h. Wang and H.-C. Liao, "Data mining for adaptive learning in a tesl-based e-learning system," Expert Systems with Applications, vol. 38, no. 6, pp. 6480–6485, 2011.

[13] N. Thai-Nghe, L. Drumond, T. Horvath, L. Schmidt-Thieme ´ et al., "Multi-relational factorization models for predicting student performance," in Proc. of the KDD Workshop on Knowledge Discovery in Educational Data. Citeseer, 2011.

[14] A. Toscher and M. Jahrer, "Collaborative filtering applied to educational data mining," KDD cup, 2010.

[15] R. Bekele and W. Menzel, "A Bayesian approach to predict performance of a student (bapps): A case with ethiopian students," algorithms, vol. 22, no. 23, p. 24, 2005.