ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 11, April 2018

TRACKING AND PREDICTING STUDENT PERFORMANCE IN GRADUATION PROGRAM USING TIME SERIES FORECASTING

Supraja S
B.E. Computer Science
Sri Venkateswara College of Engineering, Chennai,
Tamil Nadu, India.
suprajasr17@gmail.com

Poorani S
Assistant Professor – Computer Science
Sri Venkateswara College of Engineering, Chennai,
Tamil Nadu, India.
spoorani@svce.ac.in

***Abstract*— Making higher education affordable has a notable impact on assuring the nation's economic prosperity and represents a central focus of the government when making education policies [1]. Yet student loan debt in India has increased, exceeding Indians' 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 India have on-time graduation rates at or above 50 percent for their full-time students. Accurately predicting students' future performance based on their ongoing scholastic records is crucial for effectively carrying out necessary interventions to ensure students' on-time and satisfactory graduation. In this article, a novel machine learning method for predicting students' performance is described.**

Keywords*— *Time Series Forecasting, Gaussian process, weka, Performance, Prediction

1. INTRODUCTION

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 [2]-[5] or completing courses in classroom settings or in Massive Open Online Courses (MOOC) platforms [6][7]. 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.

First, students can differ tremendously in terms of background as well as their chosen area, resulting in different selected courses as well as course sequences. Since predicting student performance in a particular course relies on the student past performance in other courses, Time Series Forecasting can be used to effectively predict students' future performance based on their past or on-going academic records irrespective of their background.

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 "Data Structures" for predicting his/her grade in the course "Algorithms", the student's grade in the course "Chemistry Lab" may have much weaker predictive power. Therefore instead of taking all the courses into consideration, students' GPA in each semester is taken into consideration for predicting their final semester GPA.

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

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 11, April 2018

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 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. This work focuses on predicting students' performance by taking students' evolving progress into consideration.

II RELATED WORK

Student retention is an important issue in education. While intervention programs can improve retention rates, such programs need prior knowledge of students' performance (Yadav et al., 2012). That is where performance prediction becomes important. The usage of machine learning to predict either the student performance or the student dropout is a commonly found subject in academic literature. Dropout prediction in virtual learning, or e-learning is a particularly common focus in such studies, due to both high dropout rates and easily available data (Kalles and Pierrakeas, 2006). Areas outside of virtual learning are also common contexts where dropout or performance predictions are used for research. The purpose of the research of these studies varies. In some of them, the aim is to find the best method for prediction. In others, the aim is simply to evaluate whether machine learning is a viable approach for predicting student dropout or performance.

One study evaluating the effectiveness of machine learning for dropout prediction was done at the Eindhoven University of Technology (Dekker et al., 2009). Basic methodology was to build multiple prediction models using different machine learning methods, such as CART, BayesNet, and Logit. Then, prediction results of different models were compared in terms of their effectiveness. Most successful model was built by using the J48 classifier. (Dekker et al., 2009).

A similar study was made by researchers from three different universities in India (Yadav et al., 2012). A data set of university students was analyzed by different algorithms, after which precision and recall values of the predictions were compared. The ADT decision tree model provided the most accurate results (Yadav et al., 2012).

However, predicting student performance instead of student dropouts is more related with this thesis, and there are examples of such studies as well. One of these studies, made in the Hellenic Open University, analyzed the usage of machine learning in distance education (Kalles and Pierrakeas, 2006). Genetic algorithms and decision trees were used to build a predictive model, and the results were compared in terms of accuracy. Most accurate results were provided by the GATREE (genetically evolved decision trees) model (Kalles and Pierrakeas, 2006).

The last study reviewed here was also about performance prediction. It was done at the University of Minho, Portugal (Cortez and Silva, 2008). The data set contained information about whether the student had passed the exam in the subjects of math and Portuguese language. Decision trees, random forest, neural networks, and support vector machines were used (Cortez and Silva, 2008). These methods were compared in terms of accuracy. Another comparison was made between a data set that included the past exam results and the one that did not. Inclusion of the past grades resulted in an improved performance.

The pattern is similar in most of these studies. First, different algorithms are applied to a data set to build prediction models. Then, predictions made by these models are compared using common evaluation criteria, such as accuracy, precision, and recall. Feature selection is also a commonly compared criterion. However, what these studies are missing is considering the evolving performance of students over a particular period of time and avoiding large junk of data sets. This is the part where this thesis can introduce a new approach. By comparing the effectiveness of different processes used in machine learning, this thesis can provide insight into the more efficient way to improve predictions in student performance.

III METHODOLOGY

A common definition of machine learning is (Mitchell, 1997):

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."

Basically, machine learning is the ability of a computer to learn from experience (Mitchell, 1997). Experience is usually given in the form of input data. Looking at this data, the computer can find dependencies in the data that are too complex for a human to form. Machine learning can be used to reveal a hidden class structure in an unstructured data, or it can be used to find dependencies in a structured data to make predictions. Latter is the main focus of the thesis.

Predictive analytics is the act of predicting future events and behaviors present in previously unseen data, using a model built from similar past data (Nyce, 2007; Shmueli, 2011). It

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 11, April 2018

has a wide range of applications in different fields, such as finance, education, healthcare, and law (Sas, 2017). The method of application in all these fields is similar. Using previously collected data, a machine-learning algorithm finds the relations between different properties of the data. The resulting model is able to predict one of the properties of future data based on properties (Eckerson, 2007).

Making predictions about the future is called extrapolation in the classical statistical handling of Time Series Data. More modern fields focus on the topic and refer to it as Time Series Forecasting. Forecasting involves taking models fit on historical data and using them to predict future observations. Descriptive models can borrow for the future (i.e to smooth or remove noise), they only seek to best describe the data. An important distinction in forecasting is that the future is completely unavailable and must only be estimated from what has already happened. The purpose of time series analysis is generally twofold: to understand or model the stochastic mechanisms that gives rise to an observed series and to predict or forecast the future values of a series based on the history of that series. Time series data has a natural temporal ordering - this differs from typical data mining/machine learning applications where each data point is an independent example of the concept to be learned, and the ordering of data points within a data set does not matter. Examples of time series applications include: capacity planning, inventory replenishment, sales forecasting and future staffing levels.

Weka (>= 3.7.3) now has a dedicated time series analysis environment that allows forecasting models to be developed, evaluated and visualized. This environment takes the form of a plugin tab in Weka's graphical "Explorer" user interface and can be installed via the package manager. Weka's time series framework takes a machine learning/data mining approach to modeling time series by transforming the data into a form that standard propositional learning algorithms can process. It does this by removing the temporal ordering of individual input examples by encoding the time dependency via additional input fields. These fields are sometimes referred to as "lagged" variables. Various other fields are also computed automatically to allow the algorithms to model trends and seasonality. After the data has been transformed, any of Weka's regression algorithms can be applied to learn a model. An obvious choice is to apply multiple linear regression, but any method capable of predicting a continuous target can be applied - including powerful non-linear methods such as support vector machines for regression and model trees (decision trees with linear regression functions at the leaves). This approach to time series analysis and forecasting is often more powerful and more flexible that classical statistical techniques such as ARMA and ARIMA [8]. In this work, the performance of students in their final semester is predicted based on their on-going academic records by mentioning the time period between each and every semester in the forecaster.

Date	Marks	
2014-06-01	84	- Sem 1
2015-01-02	73	- Sem 2
2015-06-03	55	- Sem 3
2016-01-04	62	- Sem 4
2016-06-05	97	- Sem 5
2017-01-06	67	- Sem 6
2017-06-06	72	- Sem 7

III SYSTEM ANALYSIS

A. UNDERSTANDING THE OBJECTIVE

The first step in developing a project is to understand the objective which involves an understanding of the intent and essentials of a system. This comprehension is used as a problem description and a preparatory system to accomplish the expectations. The objective of our project is neither to build a system that makes billions nor to waste billions too. But the objective is to develop a system that predicts the performance of the students based on the co-relations between previous academic records and help the students and institutions to increase the pass percentage.

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 11, April 2018

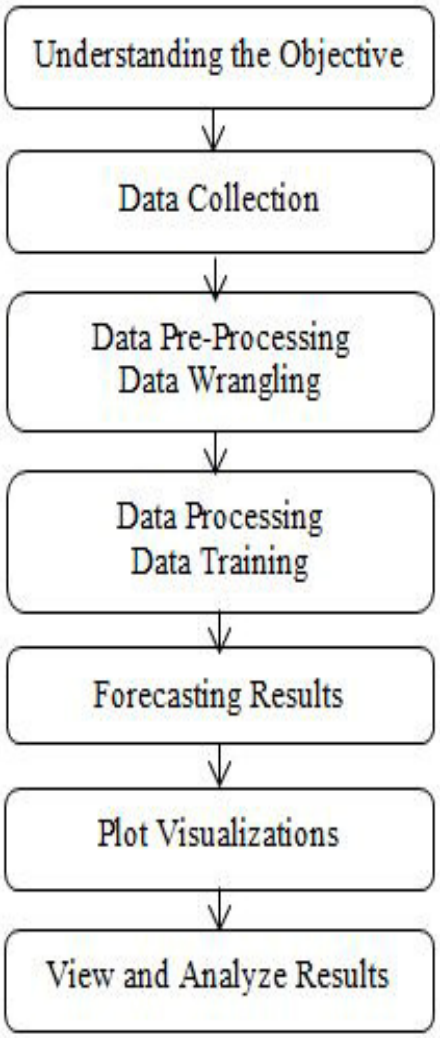


Figure 1. System Architecture [8]

B. DATA COLLECTION

Once the understanding of the objective is over, the next step is to collect the data. Data collection involves the understanding of initial observations of the data to identify the useful subsets from hypotheses of the hidden information. In this case the previous semester marks are collected as a input from the user and their evolving states for each semester if identified.

C. DATA PRE-PROCESSING

The data pre-processing stage involves all the activities to prepare the final data from the preparatory raw information. The data preparation tasks can be performed several times as there is no specific order. These tasks include the selection of a record, table, attribute and cleaning of data for modeling tools. In our methodology, the input data will be converted into a combined value vector for getting the final GPA.

D. DATA PROCESSING

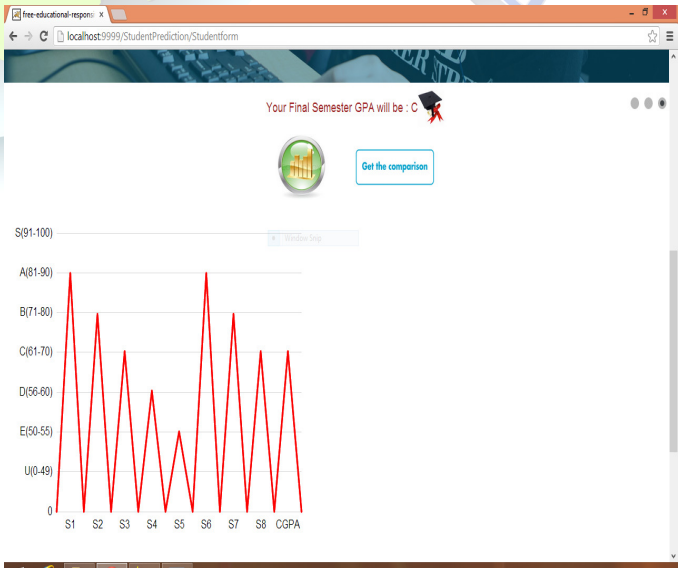
To process the data forecast() function is used. In technical analysis institutions use the auto regressive and moving average models to forecast the performance. Major steps involved here are identification, parameter estimation and forecasting. These steps are repeated until an appropriate model is identified for prediction.

E. FORECASTING RESULTS

The process of making predictions of the future by relying upon the past and present data is known as forecasting. Prediction also offers a significant standard for organizations that have a long-term perception of actions. We use 'forecast' package for predicting the furture GPA based on the analysis of past GPAs. This 'forecast' package provides a number of forecasting functions for displaying the time series predictions along with the exponential smoothing and space models.

F. PLOT VISUALIAZATIONS

Data visualization is a graphical representation of the numerical data. In our methodology, after forecasting the future GPA we visualize the results for taking necessary interventions in-terms of line charts, candlesticks charts, bar charts, and histograms. Here x-axis shows the time period in-terms of year/months/days and y-axis shows the performance of a student.



ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 11, April 2018

Figure 2. Graphical Representation of Performance

G. VIEW AND ANALYSE RESULTS

Once after plotting the results in-terms of visualizations we can find out the correlations to get the short-term predictions.

IV DISCUSSION AND CONCLUSION

The success of machine learning in predicting student performance relies on the good use of the data and machine learning algorithms. Selecting the right machine learning method for the right problem is necessary to achieve the best results. However, the algorithm alone cannot provide the best prediction results. The process of considering evolving progress in data for machine learning, is also an important factor in getting the best prediction results. The aim of this thesis was to predict the performance of the students in graduation level by using Time Series Forecasting to take their evolving states into consideration and to help the institutions in India and all around the world to achieve more pass percentage and on-time graduation. By employing this method future grades can be predicted and special attention can be given to those who are categorized under slow learners to increase the on-time graduation percentage of the institutions.

References

[1] The White House, "Making college affordable," 2016 [Online].Available:https://www.whitehouse.gov/issues/education/higher-education/makingcollege-affordable

[2] 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.* New York, NY, USA: Springer, 2006, pp. 164–175.

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

[4] H.-F. Yu *et al.*, "Feature engineering and classifier ensemble for KDD Cup 2010," in *Proc. KDD Cup 2010 Workshop*, 2010, pp. 1–16.

[5] 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," *J. Mach. Learn. Res. W & CP*, pp. 1–16, 2010.

[6] Y. Meier, J. Xu, O. Atan, and M. van der Schaar, "Personalized grade prediction: A data mining approach," in *Proc. 2015 IEEE Int. Conf. Data Mining*, 2015, pp. 907–912.

[7] C. G. Brinton and M. Chiang, "MOOC performance prediction via clickstream data and social learning networks," in *Proc. 2015 IEEE Conf. Comput. Commun (INFOCOM)*, 2015, pp. 2299–2307.

[8] Mahantesh Angadi and Amogh Kulkarni, "Time Series Data Analysis for Stock Market Prediction using Data Mining Techniques with R," in *International Journal of Advanced Research in Computer Science*, 2015, pp. 104-108.