Journal of Critical Reviews

ISSN- 2394-5125 Vol 7, Issue 3, 2020

Review Article

COMPARATIVE ANALYSIS OF ENSEMBLE ALGORITHMS'PREDICTION ACCURACIES IN EDUCATION DATA MINING

EU JIN PHUA[1], NOWSHATH KADHAR BATCHA[2]

[1]School of Computing, Engineering and Technology Asia Pacific University of Technology and Innovation, Malaysia.
[2]School of Computing, Engineering and Technology Asia Pacific University of Technology and Innovation, Malaysia.

Received: 07.11.2019 *Revised: 09.12.2019* *Accepted: 11.01.2020*

Abstract
Data from the Education domain is increasing exponentially. Education Data Mining is an evolving discipline focused on developing methods that could use the massive data collected from educational settings to discover new insights about how people learn in the context of such settings. This work is aimed at developing a predictive model that applies machine learning techniques to predict students' grades of modules from past results. It can help students to improve their performance based on the predicted scores to understand their academic status approximately and enable their instructors to identify students who might need additional assistance in the respective modules. This is possible by exploring the correlation between the grades of different modules and estimating the target subject grade. This study uses a synthetic data set and evaluates the performance of different base algorithms such as Linear Regression, K-Nearest Neighbor and Decision Table. Hybrid algorithms which are a combination of multiple base algorithms known as Ensemble algorithms were also developed and evaluated. The Ensemble algorithms used for this study include Stacking, Bagging and Random Forest. The results obtained show that the Ensemble algorithms perform better than the base algorithms when making student grade predictions

Keywords: Affinity Analysis, Education Data Mining, Ensemble Algorithms.

DOI: http://dx.doi.org/10.31838/jcr.07.03.06

INTRODUCTION

There is no grade prediction system that is reliable and readily available currently to predict student grades and provide recommendations for optional modules based on individual student's performances in previous modules. There are a few universities including the Asia Pacific University of Technology and Innovation (APU) that offer optional modules to its students. Computer science and software engineering students in APU have around one optional module per semester and must choose these modules from an option of two to four modules. Students generally want to a module or modules that they can score the best possible grade. However, most students are not able to make up their minds on their choice of optional modules as they are unsure of which modules suit them the best. Besides that, lecturers also want to know the grade prediction of each individual student. This allows lecturers to target and focus on a specific group of students that have the potential to score better grades and raise the overall grades of a specific module.

METHOD AND MATERIALS

Algorithm Comparisons

There are several algorithms that can be used to develop a student grade prediction system. Reference [1] discussed the suitable data analysis methods to analyze student grades such as Decision Tree, Neural Network, Naïve Bayes, and K-Nearest Neighbor. The Neural Network algorithm has the highest accuracy followed by Decision Tree, K-Nearest Neighbor and finally Naïve Bayes which has the lowest accuracy. This paper concludes that the Neural Network and Decision Tree are two of the most suitable and frequently used methods to predict student performance.

Reference [2] also discussed the use of different algorithms. The paper used the Decision Tree Classifier (ID3), K-Nearest Neighbor (k-NN), Bayesian Classifier and Rule Induction algorithms to predict student grades. The paper summarized the percentage accuracy of these algorithms with Rule Induction as the most accurate algorithm, followed by ID3, k-NN and Naïve Bayes. The Rule induction algorithm can predict, on average, more than 96% of student results before the final exam. The grades classification used in this paper is A, B+, B, C+, C and F. The grades which were the easiest to predict were grades F and C while the hardest grade to predict correctly was B+.

Different algorithms may vary greatly in prediction accuracy. Reference [3] discussed the comparisons in accuracy between various Data Mining Algorithms which consists of the Multilayer Perceptron (MLP), Iterative Dichotomiser 3 (ID3), J48, REP Tree, NB Tree, Simple Cart and Decision Table algorithms. All the algorithms were tested with a small dataset of 165 student records using the Waikato Environment for Knowledge Analysis (WEKA). Based on the results, MLP, a Neural Network based algorithm had the highest prediction accuracy at 74.8%, followed by ID3, NB Tree, REP Tree, Simple Cart, Decision Table and J48.

Reference [4] further discussed the use and comparisons of different algorithms. The paper compared the User-based Collaborative Filtering (UBCF), Matrix Factorization (MF), and Restricted Boltzmann Machines (RBM) algorithms. The UBCF algorithm make predictions by comparing students with similar grades. MF is a decomposition of a matrix into two or more matrices. It is used to discover hidden patterns and predict missing values of a matrix. MF consists of Singular Value Decomposition (SVD) and Non-Negative Matrix Factorization (NMF). RBM is an algorithm used to find structural patterns in a dataset. The prediction accuracy is measured by calculating the Root Mean Squared Error (RMSE), Mean Squared Error (MSE) and Mean Absolute Error (MAE) of each algorithm. The paper concluded that the RBM algorithm outperforms the other algorithms with a smaller probability of error, followed by NMF, SVD and UBCF.

Advanced Algorithms

Algorithms that make predictions based on a subset of data were able to achieve better prediction accuracy. Reference [5] used Linear Regression and Matrix Factorization to predict student grades. The methods used were Course-Specific Regression (CSR) and Student-Specific Regression (SSR). CSR is a method in which only the results of students who took a specific course is considered when making grade predictions. For example, the grade predictions of a computer science student will be made only based on previous computer science students instead of all students who may have taken similar modules. SSR is a method used to make grade predictions based on results of students who took the same optional modules within the same course. SSR caters to a variety of module combinations for flexible degree programs with many optional modules. The paper concluded that using a course specific subset (CSR) of a dataset leads to more accurate predictions. However, the prediction accuracy of a student-course specific approach (SSR) is limited by the diverse combination of modules.

There are also algorithms that are able to make more accurate predictions although the instances in a dataset are scattered. Reference [6] proposed a hybrid algorithm which consists of Regression Setting, Classification Setting and Confidence-Learning Prediction algorithms to predict student grades. These algorithms, when combined and used together, are capable of detecting patterns and trends in the data which is missed by Linear and Logistic Regression whereby a single parameter per performance criteria has to fit all students. The paper concluded that the proposed algorithm increases in accuracy (decrease in average absolute percentage error) when more data is obtained over time.

Advanced algorithms which are derived from single model algorithms are also able to achieve higher prediction accuracy. Reference [7] proposed the Ensemble-based Progressive Prediction (EPP) algorithm, which uses the Exponentially Weighted Average Forecaster (EWAF) Ensemble learning technique. The EPP algorithm enables progressive prediction as more student data is obtained and scales easily with the increasing modules taken by students. Therefore, the EPP algorithm is better for designing prediction systems for progressively expanding datasets compared to the conventional EWAF algorithm. The paper also benchmarked the EPP algorithm against the linear regression, logistic regression, random forest and k-Nearest Neighbors (kNN) algorithms. The Mean Square Error for the EPP algorithm with Time is the lowest, followed by Random Forest, kNN, Linear Regression and Logistic Regression. This indicated that EPP was the most accurate when making predictions.

Reference [8] discussed the use of advanced algorithms created from combining multiple single model algorithms. The paper used an Ensemble model (combination of multiple models) to increase the accuracy of grade predictions compared to single model-based techniques. The paper proposed an Ensemble classifier framework which eliminates random misclassified data using the Ensemble Filtering (E) algorithm to improve prediction accuracy. E consists of the Decision Tree (J48), Random Forest and Naïve Bayes algorithms. The number of Mathematics and Portuguese student grade records used were 395 and 649 respectively. Ensemble Filtering had the highest prediction accuracy, followed by Bagging and Decision Tree. The paper concluded that using Ensemble filters can greatly improve prediction accuracy compared to using single filters.

Research Methods

Synthetic Data was used to generate student grades and results due to the fact that real student results are classified as sensitive information and most universities are not willing to provide student academic results for academic research. This is because sharing such data to any third party will often risk disclosing private information of students from the university which can be misused by adversaries. Besides that, synthetic data can be generated to be close to real world data. Synthetic data can also be generated in larger amounts when necessary compared to the limited amount of real-world data that must be collected over an extensive period of time [9].

Mockaroo is an online tool that let users generate synthetic datasets and was used to generate the student scores datasets used in this study. Datasets are generated using a data schema which contains data structures predefined by the user. The data schema consists of Field Name and Field Type components. The Field Name is the attribute (module) name in the dataset and the Field Type is a numerical value (student score) associated to the Field Name (module). The Field Type is set to follow a Normal Distribution when generating student scores. The number of rows (instances) required can also be defined by the user. The dataset can then be generated based on the Field Names and Field Types in the Data Schema. All datasets generated in Mockaroo can be downloaded for testing and modelling [9].

The Synthetic Dataset was generated using Mockaroo based on a module called "Astronomy 162", which contains the summary of statistics of student scores from Ohio State University [10]. The overall student scoring statistics has a mean of 68.18% and a standard deviation of 15.92%. The resultant Grade Curve was symmetrical and bell-shaped, indicating that it has a Normal Distribution. Therefore, a Normal Distribution curve with the mean of 68.18% and a standard deviation of 15.92% was taken as a baseline to generate synthetic student results for this study. A total of two 100 instance datasets and two 1000 instance datasets were generated for this study. TABLE I shows the description of each instance (module) in the synthetic dataset. In order to create variation and simulate dependencies between modules, the latter five modules (Module D, Module E, Module F, Module G and Module H) were generated with dependencies on the first three modules (Module A, Module B and Module C). The first three algorithms were generated based on the baseline Normal Distribution curve as discussed above.

Table I: Data Field Description

Field Name	Description
Module A	Random instances of Module A generated with baseline Normal Distribution curve
Module B	Random instances of Module B generated with baseline Normal Distribution curve
Module C	Random instances of Module C module generated with baseline Normal Distribution curve
Module D	Random instances of Module D generated from a Normal Distribution curve with standard deviation of 5 which is then added with the instance from Module B
Module E	Random instances of Module E generated from a Normal Distribution curve with standard deviation of 10 which is then added with the instance from Module A
Module F	Random instances of Module F generated from a Normal Distribution curve with standard deviation of 3 which is then added with the instance from Module A
Module G	Random instances of Module G generated from a Normal Distribution curve with standard deviation of 5 which is then added with the instance from the Module D
Module H	Random instances of Module H generated from a Normal Distribution curve with standard deviation of 10 which is then added with the instance from the Module A

Module A	Module B	Module C	Module D	Module E	Module F	Module G	Module H
93	66	60	64	88	96	61	90
60	41	71	35	62	60	42	58
81	41	70	40	91	78	45	78
49	50	63	50	56	47	48	49
55	77	64	70	62	58	74	60
87	60	82	58	92	87	53	99
78	99	71	89	81	81	95	81
74	66	54	68	71	78	76	70
81	60	77	53	95	79	56	75
68	72	69	69	80	65	71	84
78	73	80	75	70	79	76	82
64	80	69	69	60	62	64	69
81	77	82	82	86	84	79	78
90	87	70	88	87	90	86	93
62	77	83	82	56	62	83	51

Fig. 1: Part of sample dataset generated via Mockaroo

Fig. 1 shows part of sample dataset generated via Mockaroo. Four datasets were generated, one 100 instance dataset and one 1000 instance dataset for training and the other 100 instance dataset and 1000 instance dataset for testing the prediction algorithms.

Analysis of Data

The Waikato Environment for Knowledge Analysis (WEKA)is an open source software that was used to test and compare the prediction accuracies of the algorithms. WEKA consists of a collection of machine learning algorithms used in data mining. It provides a test bench to test such algorithms using any dataset uploaded and provides data pre-processing and visualization tools that allow users to compare prediction results of multiple algorithms. It is also suitable for the development of new machine learning algorithms (Ensemble algorithms) discussed below [11].

The base algorithms chosen are Linear Regression, Decision Table and K-Nearest Neighbor. The Ensemble algorithms consist of Random Forest, Stacking and Bagging. Random Forest will be used as a baseline comparison for the Ensemble algorithms.

A Random Forest consists of a collection of Decision Trees as the base classifier. Each tree provides a class prediction and the class with the most votes is used as the final prediction [11].

Stacking combines the predictions of several base algorithms. The predictions of the base algorithms are then combined by a meta classifier to produce a final prediction. The base algorithms used are Linear Regression and Decision table and the meta classifier is Linear Regression [11].

Bagging is a statistical estimation technique in which a statistical quantity such as a mean is estimated from multiple random samples (subsets) of the data. Multiple random samples of the training data are drawn and replaced to train a base model. The results of the data subsets of these estimates are then averaged to achieve a lower Mean Absolute Error (MAE). The base model used is Linear Regression [11].

The algorithms will be used to predict Module Hwhich is the last module in the dataset. The algorithms are first trained and tested using the smaller pair of 100 instance datasets (one for training and one for testing). The results for the smaller test set are recorded. The same algorithms will then be tested using the larger pair of 1000 instance datasets (one for training and one for testing). The results for the larger test set are recorded and comparisons will be made for the two test sets.

RESULTS AND DISCUSSION

Table II: Mean Absolute Error (MAE) of Algorithms

Algorithm	Mean Absolute Error (MAE)	
	100 instances test set	1000 instances test set
Linear Regression	8.6513	7.5990
Decision Table	**8.2194**	7.8453
K-Nearest Neighbor	11.570	10.672
Random Forest	8.5823	8.1037
Stacking	8.4092	**7.5989**
Bagging	8.4433	7.6055

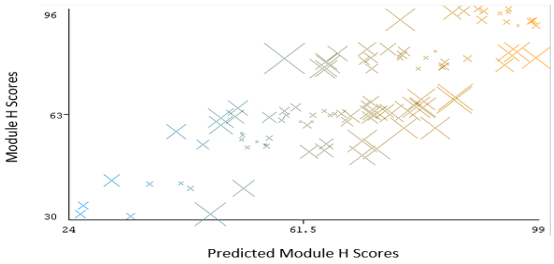

Fig. 2: Grade Predictions for Module H from the 100-instance test set using Decision Table

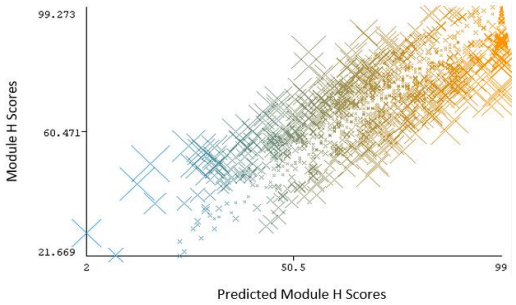

Fig. 3: Grade Predictions for Module H from the 1000-instance test set using Stacking

TABLE II shows the Mean Absolute Error (MAE) of all the algorithms tested using the 100 instance (small) dataset and the 1000 instance (large) dataset. Figures 2 and 3 show the respective scatter plot diagrams for the grade predictions of Module H in the 100-instance dataset using Decision Table and 1000-instance dataset using Stacking. Based on the results obtained, the Decision Table performed the best with the lowest MAE of 8.2194 for the small dataset. The three Ensemble algorithms performed marginally worse than the Decision Table. Stacking performed the best among the three Ensemble algorithms with a MAE of 8.4092.

However, the results differed when the algorithms were tested with the large dataset. This time, Stacking performed the best with the lowest overall MAE of 7.5989. It was also noted that Linear Regression outperformed Decision Table and had the lowest MAE among the base classifiers. Besides that, all algorithms performed better in the large dataset compared to the small dataset. This was expected as the information of algorithms discussed in the Method and Materials section suggested that classification and regression algorithms were able to perform better with larger datasets.

The benefits of Ensemble algorithms were only noticeable when a large dataset was used in which Stacking outperformed the base algorithms and the results obtained support the notion that Ensemble algorithms were able to perform better than individual base algorithms. The results also show that Bagging improved its prediction accuracy with large datasets as it has the third lowest MAE among all algorithms and almost outperformed Linear Regression. Therefore, the results suggest that the Ensemble algorithms may perform even better than the base algorithms if a dataset with more than 1000 instances is used for training and testing. A dataset of 10,000 instances is proposed for training and testing algorithms in the future. Besides that, more complex Ensemble algorithms which combines three to five base algorithms will be developed in the future. On top of that, an Ensemble algorithm can also be made froma combination of multiple other Ensemble algorithms, forming a more complex and robust Ensemble algorithm which is potentially more sensitive to data variations. These algorithms will be explored and tested in the future.

CONCLUSION

Based on the findings from the results and discussion, it was proven that Ensemble algorithms were able to perform better than base algorithms, provided that the dataset is large. This is because if the dataset is too small, the Ensemble algorithm will not have enough instances to make predictions which cause the predictions to be skewed. Algorithms may also increase in prediction accuracy as verified by the results obtained. An attempt to improve the accuracy of the Ensemble algorithms will be made. More complex Ensemble algorithms will be developed and tested in the future. These algorithms will consist of at least three base algorithms and may be a combination of multiple other Ensemble algorithms. The potential of Ensemble algorithms to outperform base algorithms were verified through testing and benchmarking. These Ensemble algorithms performed significantly better with large datasets. Therefore, the Ensemble algorithms used in this study will be trained and tested again with a larger dataset of 10,000 instances. These Ensemble algorithms will then be used as a benchmark/baseline for comparisons with more complex algorithms that will be developed in future studies. The results will be evaluated to determine the increase in accuracy of the Ensemble algorithms from the base single model algorithms. The proposed benchmarking technique is to calculate the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) of each algorithm.

REFERENCES

1. Shahiri, Husain and Rashid (2015). A Review on Predicting Student's Performance using Data Mining Techniques. In The Third Information Systems International Conference Indonesia, Monday 2nd to Wednesday 4th November 2015 Indonesia: Elsevier, pp.414-422
2. Majeed and Jujeno (2016). Grade Prediction Using Supervised Machine Learning Techniques. In The 4th Global Summit on Education GSE 2016 Malaysia, Monday 14th to Tuesday 15th March 2016 Malaysia: WorldConferences, pp. 222-234
3. Ruby and David (2014). Predicting the Performance of Students in Higher Education Using Data Mining Classification Algorithms - A Case Study. International Journal for Research in Applied Science & Engineering Technology (IJRASET). 2(11) pp. 173-180
4. Iqbal, Qadir, Mian and Kamiran (2017). Machine Learning Based Student Grade Prediction: A Case Study. Cornell University Library [Online] 1(8) pp. 1-22. Available from: https://arxiv.org/pdf/1708.08744.pdf [Accessed 24/05/2018]
5. Polyzou and Karypis (2016). Grade Prediction with Course and Student Specific Models. In The Pacific-Asia Conference on Knowledge Discovery and Data Mining 2016 New Zealand, Tuesday 19th to Friday 22nd April 2016 New Zealand: Springer International Publishing, pp. 89-101
6. Meier, Xu, Atan and Schaar (2016) Predicting Grades. IEEE Transactions on Signal Processing. 64(4) pp. 959-972
7. Xu, Moon and Schaar (2017) A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs. IEEE Journal of Selected Topics in Signal Processing. 11(5) pp. 742-753
8. Satyanarayana and Nuckowski (2016) Data Mining using Ensemble Classifiers for Improved Prediction of Student Academic Performance. In ASEE Mid-Atlantic Section Spring 2016 Conference Washington, Friday 8th to Saturday 9th April 2016 Washington: CUNY Academic Works, pp. 1-7
9. Matthews (2016). How Cooperative Game Theory can be utilised to enhance marketing analytics attribution. Ireland: National College of Ireland
10. Pogge (2005). A Brief Note about Grade Statistics or How the Curve is Computed. Astronomy 162 [Online]. Available from: http://www.astronomy.ohio-state.edu/~pogge/Ast162/Quizzes /curve.html [Accessed 9/05/2019]
11. Weka (2018). Weka 3: Data Mining Software in Java. Machine Learning Group at the University of Waikato. [Online]. Available from: https://www.cs.waikato.ac.nz/ml/weka/ [Accessed 9/05/2019]
12. Shrivastava, S., Jeyanthi, P.M. and Singh, S., 2020. Failure prediction of Indian Banks using SMOTE, Lasso regression, bagging and boosting. Cogent Economics & Finance, 8(1), p.1729569.
13. Aditi Chaturvedi, Rangeel Singh Raina, Vijay Thawani, Harish Chaturvedi, Deepak Parihar. "Super TB: Another Manmade Disaster." *Systematic Reviews in Pharmacy* 3.1 (2012), 37-41. Print. doi:10.4103/0975-8453.107140