ISSN: 2278-4632
Vol-11 Issue-01 2021

CLASSIFYING STUDENTS PERFORMANCE USING GRADIENT BOOSTING ALGORITHM TECHNIQUE

Mr. SK.MAHABOOB BASHA[1],
K.S.S.S.PREETHI[2], M.POOJA3, G.MEGHANA[4]

[1]Associate Professor, Dept. of INFORMATION TECHNOLOGY, NRI INSTITUTE OF TECHNOLOGY , A.P., India.
[2,3,4]Student, B.Tech (IT), NRI INSTITUTE OF TECHNOLOGY , A.P., India.

Abstract — . In recent years due to the rapid development of technology the amount of data has been growing tremendously in all areas. The need of discovering novel and most useful information from these large amounts of data has also grown. With the advent of data mining, different mining techniques have been applied in different application domains, such as, Education, banking, retail sales, bioinformatics, and Telecommunications. The objective of our work is two-fold. First, inorder to overcome this limitation, we explore if poorly per-forming students can be more accurately predicted by formulating the problem as binary classification. Second, in orderto gain insights as to which are the factors that can leadto poor performance, we engineered a number of human-interpretable features that quantify these factors. Thesefeatures were derived from the students' grades from theUniversity of Minnesota, an undergraduate public institution.

INTRODUCTION

Education data processing helps in predicting students’ performance so as to recommend improvements in academics. The past several decades have witnessed a rapid climb within the use of knowledge and knowledge mining as a tool bywhich academic institutions extract useful unknown information in the student result repositories in order to improve students’ learning processes .The main objective of this project is prediction of student’s performance based on random forest classification technique using toolssuch as WEKA ,ORANGE and scikit-learn libraries in python. Higher educational institutions constantly try to improve the retention and success of their enrolled students. Accord- ing to the US National Center for Education Statistics [8], 60% of undergraduate students on four-year degrees will not graduate at the same institution where they started within the rest six years. At the same time, 30% of college fresh- men drop out after their rest year of college. As a result, colleges look for ways to serve students more efficiently and ectively. This is where data mining is introduced to pro- vide some solutions to these problems. Educational data mining and learning analytics have been developed to pro- vide tools for supporting the learning process, like monitor and measure student progress, but also, predict success orguide intervention strategies. Many prediction models available with a difference in approach to student performance were reported by the researcher, but there is no certainty that there are any predictors who can accurately determinewhether a studentwill be anacademic genius, a drop out, or an average performer. The higher education institutions use automated computer programs/tools developed with different technologies to predict the trades in the college. With the potential techniques in Data Mining and with the growth of technologies to handle huge databases, the predictive technologies have started growing tremendously. The academic research in Data Mining also contributed a lot to predictive technologies. The prediction of academic performance is regarded as a challenging task of temporal data prediction. Data analysis is one way of predicting increase or decrease of future academic performance.

LITERATURE SURVEY

1.Support-vector networks. Machine learning

The support-vector network is a new learning machine for two-group classification problems. The machine conceptually implements the following idea: input vectors are non-linearly mapped to a very highdimension feature space. In this feature space a linear decision surface is constructed. Special properties of the decision surface ensures high generalization ability of the learning machine. The idea behind the support-vector network was previously implemented for the restricted case where the training data can be separated without errors. We here extend this result to non-separable training data. High generalization ability of support-vector networks utilizing polynomial input transformations is demonstrated. We also compare the performance of the support-vector network to various classical learning algorithms that all took part in a benchmark study of Optical Character Recognition.

2.Greedy function approximation: a gradient boosting machine.

Function estimation/approximation is viewed from the perspective of numerical optimization in function space, rather than parameter space. A connection is made between stagewise additive expansions and steepest-descent minimization. A general gradient descent “boosting” paradigm is developed for additive expansions based on any fitting criterion.Specific algorithms are presented for least-squares, least absolute deviation, and Huber-M loss functions for regression, and multiclass logistic likelihood for classification. Special enhancements are derived for the particular case where the individual additive components are regression trees, and tools for interpreting such "TreeBoost" models are presented. Gradient boosting of regression trees produces competitive, highly robust, interpretable procedures for both regression and classification, especially appropriate for mining less than clean data. Connections between this approach and the boosting methods of Freund and Shapire and Friedman, Hastie and Tibshirani are discussed.

PROPOSED SYSTEM

Most of the existing approaches focus on identifying stu- dents at risk who could benet from further assistance in order to successfully complete a course or activity. A fun- damental task in this process is to actually predict the stu dent's performance in terms of grades. While reasonable prediction accuracy has been achieved [14, 10], there is a significant weakness of the models proposed to identify the poor-performing students [18]. Usually, these models tend to be over-optimistic for the performance of students, as the majority of the students do well, or have satisfactory enough performance.

In this paper, we investigate the problem of predicting the performance of a student in the end of the semester before he/she actually takes the course. In order to focus on the poor-performing students, who are the ones that need these systems the most, the prediction problem is formulated as a classication task, where two groups of students are formed according to their course performance. We essentially iden- tify two complementary groups of students, the ones that are likely to successfully complete a course or activity, and the ones that seem to struggle. After identifying the latter group, we can provide additional resources and support to enhance their likelihood of success.

RELATED WORK

In this project author is describing concept to predict or classify student performance based on their previous academic performance. Using this paper we will concentrate more on poor performance students by extracting grade features from their past performance records. In this project we are using university dataset which contains record from A to W and we are extracting 4 features from this dataset to classify poor performing students.

1) Taking those records from dataset which has features D and F and consider as failing student and we will assign features values as 0 (Fgr) for such students.
2) Taking those records from dataset which contains dropout students and assign feature value as 1 (Wgr)

3) Taking those records from dataset which has grades lower than expected and assign feature value as 2 (RelF)
4) Taking those records from dataset which has grade value lower than expectation and he is having difficulty in study course and assign value as 3 (RelCF)
5) Rest student we are marking with feature value as 4 which indicate student is performing well.

By using 'University ofMinnesota' grade dataset we are extracting above features and assign those values as the target or class label for this dataset. After extracting features we are applying 4 machine learning algorithms on this dataset to generate training model, later new student record will be applied on this dataset to classify that student records as good performer or poor performer and we can know the reason of poor performance such as Fgr (indicate as failing student), Wgr (indicate as dropout), ReIF (lower than expected grade) or RelCF (lower than expected grade and having course difficulty).

4 algorithms used in this paper

SVM Algorithm: Machine learning involves predicting and classifying data and to do so we employ various machine learning algorithms according to the dataset. SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes. In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification. As a simple example, for a classification task with only two features (like the image above), you can think of a hyperplane as a line that linearly separates and classifies a set of data.

Intuitively, the further from the hyperplane our data points lie, the more confident we are that they have been correctly classified. We therefore want our data points to be as far away from the hyperplane as possible, while still being on the correct side of it.

So when new testing data is added, whatever side of the hyperplane it lands will decide the class that we assign to it.

Random Forest Algorithm: its an ensemble algorithm which means internally it will use multiple classifier algorithms to build accurate classifier model. Internally this algorithm will use decision tree algorithm to generate it train model for classification.

Decision Tree Algorithm: This algorithm will build training model by arranging all similar records in the same branch of tree and continue till all records arrange in entire tree. The complete tree will be referred as classification train model.

Gradient Boosting Algorithm:Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting. Gradient boosting models are becoming popular because of their effectiveness at classifying complex datasets, and have recently been used to win many Kaggle data science competitions.

The Python machine learning library, Scikit-Learn, supports different implementations of gradient boosting classifiers, including XGBoost. By using multiple algorithms a single accurate train model will be generated. In all this algorithms Gradient Boosting give better performance.

IMPLEMENTATION

classification

Our motivation was to identify groups of students that need further assistance and guidance in order to successfully complete a course. These students could benet from informed interventions. We consider this to be a binary classification problem, where these students form one of the classes and the remaining students form the other class. We consider different ways of measuring when a student does not do well in a course to deal with the performance measurement challenges we mentioned earlier. Unsatisfactory performance can occur when the earned grade represents a performance that is bellow the student's potential. We considered the following four ways for labelling, resulting to these absolute and relative classification tasks:

1.Failing student performance, i.e., letter grades D andF (denoted as the Fgr task).
2. The letter grade W (denoted as the Wgr task). Thisrepresents the instances when the student dropped thecourse. This behavior is worrisome as it shows that either the student was not interested in the course anymore or he/she expects to perform poorly.

3. Student performance that is worse than expected, i.e.,the grade achieved is more than two letter grades lowerthan the student's GPA (denoted as the RelF task).

4. Student performance that is worse than expected whiletaking into consideration the di_culty of the course(denoted as the RelCF task). The difficulty of acourse is expressed by the average grade achieved bythe students that took the course in prior offerings. Apositive instance is when the grade achieved is morethan two letter grades lower than

the average of thestudent's GPA and the course's prior average grade.

Methods compared

In order to support students that need help to successfullycomplete a course, we will use classification techniques toidentify them from the rest of the students. The instances ofinterest will be labeled as 1, and the rest as 0. The problemcan be described as follows. We are given a set of trainingexamples that are in the form (x; y) and we want to learntheir structure. We assume that there is some unknownfunction y = f(x), that corresponds the feature vector xto a value y. In our case, y = f0; 1g. A classifier is anhypothesis about the true function f. Given unseen valuesof x, it predicts the corresponding y values.

CONCLUSION

In this paper, We proposed, to accurate identify studentsthat are at risk. These students might fail the class, dropit, or perform worst than they usually do. We extractedfeatures from historical grading data, in order to test different simple and sophisticated classification methods based onbig data approaches. The best performing methods are theGradient Boosting and Random Forest classifiers, based onAUC and F1 score metrics. We also got interesting findings that can explain the student performance.

References

[1] L. Breiman. Random forests. Machine learning, 45(1):5{32, 2001.

[2] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. Classification and regression trees. CRC press, 1984.

[3] C. Cortes and V. Vapnik. Support-vector networks. Machine learning, 20(3):273{297, 1995.

[4] J. H. Friedman. Greedy function approximation: a gradient boosting machine. Annals of statistics, pages 1189{1232, 2001.

[5] Z. Iqbal, J. Qadir, A. N. Mian, and F. Kamiran. Machine learning based student grade prediction: A case study. arXiv preprint arXiv:1708.08744, 2017.

[6] J. E. Knowles. Of needles and haystacks: Building an accurate statewide dropout early warning system in wisconsin. Journal of Educational Data Mining, 7(3):18{67, 2015.

[7] S. Kotsiantis, C. Pierrakeas, and P. Pintelas. Predicting students'performance in distance learning using machine learning techniques. Applied Articial Intelligence, 18(5):411{426, 2004.

[8] J. McFarland, B. Hussar, C. de Brey, T. Snyder, X. Wang, S. Wilkinson-Flicker, S. Gebrekristos, J. Zhang, A. Rathbun, A. Barmer, et al. Undergraduate retention and graduation rates. In The Condition of Education 2017. NCES 2017-144. ERIC, 2017.

[9] B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer, and W. F. Punch. Predicting student performance: an application of data mining methods with an educational web-based system. In

Frontiers in education, 2003. FIE 2003 33rd annual, volume 1, pages T2A{13. IEEE, 2003.

[10] S. Morsy and G. Karypis. Cumulative knowledge-based regression models for next-term grade prediction. In Proceedings of the 2017 SIAM International Conference on Data Mining, pages 552{560. SIAM, 2017.

[11] E. Osmanbegovic and M. Suljic. Data mining approach for predicting student performance. Economic Review, 10(1):3{12, 2012.

[12] A. Pardo, N. Mirriahi, R. Martinez-Maldonado, J. Jovanovic, S. Dawson, and D. Gasevic. Generating actionable predictive models of academic performance. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, pages 474{478. ACM, 2016.