Which Factors Have the Greatest Impact on Student's Performance

To cite this article: Feng Li et al 2019 J. Phys.: Conf. Ser. 1288 012077

View the article online for updates and enhancements.
Which Factors Have the Greatest Impact on Student’s Performance

Feng Li, Yu Zhang, Mo Chen and Kening Gao
School of Computer Science and Engineering, Northeastern University, Shenyang, China.
Email: lifeng@mail.neu.edu.cn

Abstract. Nowadays, a great deal of educational data has been produced by E-learning system and MOOC. Educational data is important for Teaching and research. These educational data can be classified many kinds of feature, such as demographic features, social features and behavioral features. And which feature is the most import for student’s performance? In this paper, we use some common data mining technologies including Naïve Bayesian(NB), Artificial Neural Network(ANN), Support Vector Machine(SVM) and Decision Tree Classifier(DT) to predict students' performance, and try to find out the influence of characteristics on students' academic performance. From the conclusion, we can see that SVM technique outperform others, and Behavioral Features have good effect on students’ performance.

1. Introduction
Nowadays with the popularization of e-learning system and MOOC, the educational data volume is growing rapidly. Educational data play an important role in predicting students' academic performance and knowing about their learning state. Paulo Cortez and Alice Silva used data mining techniques to predict the student’s performance in secondary school first at 2008[2]. The data they used was collected by using school report and questionnaires. Elaf Abu Amrieh proposed a new student performance model with a new category of features, which called behavioral features at 2015[3]. The educational data they used was collected from different sources including traditional surveys, educational institute databases and e-learning systems.

As we know, all the educational data can hold important information including the patterns and the trends, which is important for our study about the student’s performance. Hence, how to use these resources effectively and discover the knowledge from the educational data become a problem which should be solved at present. Using traditional data mining methods can be effectively obtained these knowledge.

In this paper, we used four common data mining technologies to discover the knowledge from the educational data collected by Elaf Abu Amrieh[3], including Naïve Bayesian(NB), Artificial Neural Network(ANN), Support Vector Machine(SVM) and Decision Tree Classifier(DT). Furthermore, we try to find out the ranking of factors affecting performance, which is the most important feature for student’s performance, Gender, Teacher, or Others.

The rest of this paper consists of the following four parts: Section II elaborates the related research work on educational data mining and students’ performance. Section III elaborates the educational data and preprocessing. Section IV elaborates the results of the experiments. Section V elaborates the conclusion of the paper.

2. Related Works
It is well known that data mining technology can provide effective and referential data for the researchers. It is also an important way to convert data into information in a wide range of fields including banking, marketing, and fraud detection[1]. In recent years, data mining techniques has been applied by more and more researches in estimating the effect of factors on student’s academic performance. Among these factors, personal, family variables, socio-economic, psychological, and behavioral features ones have the greatest impact on student’s academic performance.
Cortez and Silva[2] tried to predict secondary school student performance on Mathematics and Portuguese language. They used four data mining algorithms such as Random Forest, Decision Trees, Neural Networks and Support Vector Machines. From their research, we can see that students' previous performances have a great effect on their achievements. In addition, there are also other relevant features, including: demographic (e.g. age and parent’s job), school related (e.g. number of absences and reason to choose school) and social variables (e.g. alcohol consumption).

Elaf Abu Amrieh[3] proposed a new student performance model with a new category of features in 2015, which called behavioral features. They used data mining techniques including Naïve Bayesian, Artificial Neural Networks and Decision Tree classifiers to estimate the effect of the features on student’s academic performance. The classification accuracy of the results obtained by using behavioral feature classification method is 29% higher than that of the same data set when removing behavioral characteristics. M. Ramaswami and R. Bhaskaran[4] have predicted the results of performance on high school education by using CHAID prediction model. The characteristics of secondary education achievement, type of secondary education, medium of instruction, location of school, living area are the most powerful indicators to measure the performance of students in higher secondary education. The method named CHAID is one of the classification tree algorithms. It is the name of an Automatic Interaction Detector developed for categorized variables.

Parneet Kaur et al.[5] have used classification techniques to predict the students’ academic performance and also analyze the data set of the slow learners among 152 students. These work was already analyzed and tested by five Classification algorithms including Naïve Bayes, Multilayer Perception, REPTree, SMO and J48. The results has shown that MLP performs better than other classifiers.

Arockiam et al. [6] mainly engaged in two kinds of data mining methods including K-means clustering technique and FP Tree in order to learn the knowledge from dataset of the students. We use FP Tree mining to filter the models in datasets and use K-means clustering to measure the of the students’ programming skills.

V. Ramesh et al. [7] tried to find out the factors affecting the performance students’ final exams. and an effective algorithm of data mining to predict the achievement of students in order to warn those students who are at risk in time and properly. By comparing these five classification algorithms including Naïve Bayes, Multilayer Perception, J48 and REPTreeSMO, MLP is the best one.

Nowadays, MOOC becomes more and more popular, so lots of researchers poured time and energy into the online educational data research. Yukselturk E et al.[8] studied dropout prediction through data mining methods by using an online program. They used four data mining methods based on KNN, DT, NB, NN to sort dropout students. Maria Goga et al.[9] have used Random Tree(RT) to predict performance and recommend necessary intervention strategies.

In conclusion, lots of data mining techniques have been studied in order to effectively solve the problems of education including Naïve Bayes, J48, Multilayer Perception and SMO. However, few scholars are engaged in SVM method. So in this paper, we compare SVM with other techniques in order to infer about the predictive abilities of each model. Furthermore, we also want to get the ranking of the factors which affect students’ performance.

3. Data and Preprocessing

3.1. Data Set

The data set we used in the paper is provided by Amrieh E. et al.[3], and the data set is collected from Kalboard360 E-learning system by using Experience API(XAPI)[10]. Kalboard360 is a multi-agent LMS, which has been designed to facilitate learning through the use of leading-edge technology[11]. Such system provides users with a synchronous access to educational resources from any device with Internet connection.

In this data set, it has 305 males and 175 females. And the students come from different countries, for example, it has 22 students are from Iraq, 28 students from Palestine, 172 students are from Jordan, 179 students are from Kuwait, and so on. The data come from two educational semesters: 245 student records are collected from the first semester and 235 student records are from the second semester.
The dataset also includes the attendance feature which is called “Student Absence Days”, such feature is classified into two categories based on their absence days: 289 students’ absence days under 7 and 191 students exceed 7 absence days. All the student features are given in the next table for reference.

### Table 1. Student features and description

| Feature          | Description             | Domain                  |
|------------------|-------------------------|-------------------------|
| Gender           | gender of student       | 'Female' or 'Male'      |
| NationalITy      | Nationality of student  | 'USA', 'Jordan', 'Kuwait', 'Lebanon', 'Egypt', 'KSA', 'Venezuela', 'Iran', 'Syria', 'Iraq', 'Lybia', 'Tunis', 'Morocco', 'Palestine' |
| PlaceofBirth     | Place of birth of student | 'USA', 'Jordan', 'Kuwait', 'Lebanon', 'Egypt', 'KSA', 'Venezuela', 'Iran', 'Syria', 'Iraq', 'Lybia', 'Tunis', 'Morocco', 'Palestine' |
| StageID          | student’s level         | 'HighSchool', 'MiddleSchool', 'lowerlevel' |
| GradeID          | Student’s grade         | 'G-01', 'G-02', 'G-03', 'G-04', 'G-05', 'G-06', 'G-07', 'G-08', 'G-09', 'G-10', 'G-11', 'G-12' |
| SectionID        | classroom               | 'A', 'B', 'C'           |
| Topic            | course                  | 'Spanish', 'IT', 'History', 'Math', 'Chemistry', 'Arabic', 'Biology', 'Science', 'French', 'Geology', 'Quran', 'English' |
| Semester         | school year semester    | 'First' or 'Second'    |
| Relation         | responsible for student’s education | 'father' or 'mom' |
| Raised hands     | times of raising hands  | {0,100}                |
| VisITed Resources| times of visiting a course content | {0,100}                |
| Announcemen t View| times the student checks the new announcements | {0,100}                |
| Discussion       | times of the student participates on discussion groups | {0,100}                |
| Parent Answer Survey | if the parent answered the school’s surveys | 'Yes' or 'No' |
| Parent School Satisfaction | the Degree of parents’satisfaction | 'Yes' or 'No' |
| Student Absence Days | absence days for each student | 'above-7' or 'under-7' |

In the dataset, the students are classified into three numerical intervals based on their total grade/mark: Low-Level we called ‘L’, interval includes values from 0 to 69; Middle-Level we called ‘M’, interval includes values from 70 to 89; High-Level we called ‘H’, interval includes values from 90-100. And we try to predict the grade of students’ performance.

### 3.2. Data Preprocessing

And then, we apply some preprocessing mechanisms to improve the quality of the data set. It is well known that data preprocessing is an important step in data mining process, which includes data cleaning, data reduction, data selection and data transformation. First, we use ‘L’, ‘M’ and ‘H’ to express ‘Low-Level’, ‘Middle-Level’, and ‘High-Level’ separately.

Then, we need to preprocess the category features, such as ‘Gender’, ‘NationalITy’, and so on. All of the category features can’t be expressed as a number simply, because the model also interprets the category as orderly, but the category is disordered actually. So here it can be represented by One-hot encoding. In this way, the category feature with M possible values can convert into M binary features, and only one of them is validly. Such as, ‘Gender’ need to convert into ‘Gender-F’ and ‘Gender-M’, if ‘Gender-F’ equals ‘1’, then ‘Gender-M’ equals ‘0’. 

3
4. Experiments and Results

4.1. Environment and Techniques
We ran the experiments on the computer containing 4GB of RAM, Intel(R) Core(TM) i3-2120 CPU@3.30GHz. The Operating System is Flagship Windows 7, and its type is 64 Bit. Furthermore, we use standard k-fold cross validation (k=5) to divide the dataset into training and testing. For implementation of all the classification tasks we have used Anaconda.

During our work, the classification algorithms are applied to evaluate the features that may have an effect on the performance of students. We used four Classification algorithms which are Support Vector Machine (SVM), Artificial Neural Network (ANN), Naïve Bayesian (NB) and Decision Tree classifiers (DT).

Artificial Neural Network (ANN) is an arithmetical model that is motivated by the organization and/or functional feature of biological neural networks [12]. ANN model consist of three layers: input layer, hidden layer and output layer. As a general rule an ANN is an adaptive system that adjusts its structure based on external or internal information that runs through the network during the learning process. Naïve Bayesian (NB) [13] is a very simple classification algorithm. It is called Naïve Bayesian because the idea of this algorithm is actually simple. For the given items, the algorithm calculates the conditional probability of each item. Decision Tree classifiers (DT) [13] is an automatic rule discovery technique that produces a set of branching decisions that end in a classification; it works best on nominal attributes, so numeric ones need to be split into bins. Support Vector Machine (SVM) is a kind of binary generalized linear classifier, and it is also a supervised learning system. Simply, SVM is mapping the linearly inseparable data in low-dimension space to high-dimension space and make it linearly separable.

4.2. Evaluation Criterion
As we know, Accuracy, Precision, Recall, F-Measure are four common evaluation methods in Machine Learning. They can be calculated as followed:

| Actual                  | Prediction | Negative | Positive |
|-------------------------|------------|----------|----------|
|                         |            | True Negative (TN) | False Negative (FN) |
|                         |            | False Positive (FP) | True Positive (TP) |

Based on the variables in Table 2, we can calculate Accuracy, Precision, Recall, F-Measure by the following formula.

\[
\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}
\]

\[
\text{Precision} = \frac{TP}{TP+FP} \tag{2}
\]

\[
\text{Recall} = \frac{TP}{TP+FN} \tag{3}
\]

\[
\text{F-Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}
\]

Based on the formulas above, Accuracy, Precision and Recall can be understood easily. And F-Measure is the weighted average of Precision and Recall, it is considered as a good indicator of the relationship between Precision and Recall.
4.3. Results

Table 3 presents the results using four different classification algorithms (ANN, NB, DT, SVM). According to the study of Amrieh E. et al.[3], ANN algorithm outperform DT and NB, but they did not consider about SVM model.

| Algorithm Evaluation | DT    | NB    | ANN   | SVM   |
|----------------------|-------|-------|-------|-------|
| Accuracy             | 61.02 | 71.56 | 73.51 | 73.91 |
| Recall               | 61.33 | 71.56 | 73.51 | 75.98 |
| Precision            | 60.72 | 72.71 | 73.90 | 74.00 |
| F-Measure            | 60.01 | 72.02 | 72.82 | 74.50 |

We notice that SVM model outperform others. SVM gives 73.91 Accuracy, and 74.00 Precision, these two evaluation criterion are nearly as the same as ANN. And it gives 95.98 Recall and 74.50 F-Measure, both of the values are bigger than ANN. However, our aim is not to find out the best classification model here, we aim to find out the most important features in the dataset. We applied Breadth First Search approach to check which the most important features. Table 4 shows the Top 10 Accuracy ranking of features. Furthermore, we can see that most of Behavioral Features[3] have a greater effect on students’ performance, such as Raisedhands, VisITedResources and AnnouncementsView. However, the Discussion feature is disappeared. Why? Maybe what they discussed has nothing to do with learning.

| Feature                  | Accuracy |
|--------------------------|----------|
| Raisedhands              | 0.604    |
| VisITedResources         | 0.558    |
| AnnouncementsView        | 0.523    |
| Relation                 | 0.487    |
| StudentAbsenceDays       | 0.443    |
| NationalITy              | 0.441    |
| Topic                    | 0.431    |
| PlaceofBirth             | 0.426    |
| Discussion               | 0.416    |

5. Conclusion and Future Work

In this paper, we use four data mining techniques which are Support Vector Machine (SVM), Artificial Neural Network (ANN), Naïve Bayesian (NB) and Decision Tree classifiers (DT) to predict students’ performance. We use four common evaluation methods which are Accuracy, Precision, Recall, F-Measure. The result shows that SVM outperform others, it can predict more accurately. Then, we traverse all the features, the result shows that the feature Raisedhands is the most import feature to students’ performance. It means that times the student raises hand on classroom can reflect the students’ performance very well.

Nowadays, more and more users use E-learning system and MOOC to study, so a great deal of online Behavioral Features has been produced, such as how long the students spent on watching videos, the speed of submitting jobs, and so on. Therefore, we need expand our dataset with more features. Also, Deep Learning Technique may be more effective for our experiments.

By the way, Raisedhands is a behavior under the line. Although the development of MOOC is so rapidly, but it still difficult to replace traditional education, the classroom education is still indispensable.
6. Acknowledgements
This work is supported by the National Natural Science Foundation of China (U1811261) and the National Natural Science Foundation of China (U1435216).

7. References
[1] Benoît, G. (2002). Data mining. In Annual Review of Information Science and Technology, 36, (pp.265–310).
[2] P. Cortez, and A. Silva, Using Data Mining To Predict Secondary School Student Performance, In A. Brito & J. Teixeira (eds.), EUROISIS, 2008, pp. 5-12.
[3] Amrieh, E. A., Hamtini, T., and Aljarah, I. Preprocessing and analyzing educational data set using X-API for improving student's performance. In Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on (pp. 1-5).
[4] M.Ramaswami and R.Bhaskaran, A CHAID Based Performance Prediction Model in Educational Data Mining, International Journal of Computer Science Issues Vol. 7, Issue 1, No. 1, January 2010.
[5] Kaur P, Singh M, Josan G S. Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector[J]. Procedia Computer Science, 2015, 57:500-508.
[6] Arockiam L, Charles S, Carol I, et al. Deriving Association between Urban and Rural Students Programming Skills[J]. International Journal on Computer Science & Engineering, 2010, 2(3):687-690.
[7] Ramesh V, Parkavi P, Ramar K. Predicting Student Performance: A Statistical and Data Mining Approach[J]. International Journal of Computer Applications, 2013, 63(8):35-39.
[8] Yukselturk E, Ozeke S, Yalin Kilic Turel. Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program[J]. European Journal of Open, Distance and E-Learning, 2014, 17(1):118-133.
[9] Goga M, Kuyoro S, Goga N. A Recommender for Improving the Student Academic Performance[J]. Procedia - Social and Behavioral Sciences, 2015, 180:1481-1488.
[10] Moisa, V. (2013). Adaptive Learning Management System. Journal of Mobile, Embedded and Distributed Systems, 5(2), 70-77.
[11] Amrieh, E. A, Hamtini, T, Aljarah, I.. Mining Educational Data to Predict Student’s academic Performance using Ensemble Methods. International Journal of Database Theory and Application, 2016, 9(8), 119-136.
[12] Naser S A, Zaqout I, Ghosh M A, et al. Predicting Student Performance Using Artificial Neural Network: in the Faculty of Engineering and Information Technology[J]. International Journal of Hybrid Information Technology, 2015, 8:221-228.
[13] Han J. Data Mining: Concepts and Techniques[M]. Morgan Kaufmann Publishers Inc. 2005.