Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Course performance prediction for basic courses of universities based on support vector machine

IOP Conf. Series: Journal of Physics: Conf. Series **1168** (2019) 032066 doi:10.1088/1742-6596/1168/3/032066

Course performance prediction for basic courses of universities based on support vector machine

Jiaming Chen[1], Liming Luo[1*] and Jie Song[1]

[1] College of Information Engineering, Capital Normal University, Beijing, 100048, China

[*]Corresponding author's e-mail: luolm@cnu.edu.cn

Abstract. As the popular area of educational data mining, course performance prediction is the prediction of the final performance of students in a course, assisting educators` personalized teaching and reducing the pressure of students. With proper features and algorithms specified, prediction models with high accuracy can be built. Course performance prediction for basic courses of universities requires feature selection from both subjective and objective aspects. GPA, grades of prerequisite courses, assignment scores and the inquiry count are selected as the objective features and the individual interest is specified as the subjective feature. The output of the prediction is the students` final performance divided into 5 levels. The Gaussian support vector machine, the polynomial support vector machine, BP neural network, random forest and logistic regression were employed as the classifier, with the accuracy and AP of the five algorithms compared. It is found that the Gaussian support vector machine combined with selected features can reach the optimal accuracy and AP, reaching 99%. With the Gaussian support vector machine applied, a course performance prediction model for basic courses of universities is proposed, which provides a novel method for the study on course performance prediction for basic courses of universities.

1. Introduction

As one of the most popular data mining areas, Educational Data Mining (EDM) can offer useful information used in formative evaluation to educators and a pedagogical basis for their decision when designing and altering the educational environment or method [1]. Student performance prediction is one of the key areas of EDM [2]. There are two kinds of student performance prediction methods: overall performance prediction and course performance prediction. Overall performance prediction is a composite student performance prediction. After observing for a long period, e.g. one year or four years, overall performance prediction anticipates the academic achievement of students at the end of the period. For example, Superby et al. predicted the performance and failure probability of students from three Belgian universities at the end of the first year by their personal information, course involvement and perceptions [3]; Asif et al. employed students` marks in a four-year Information Technology bachelor programme of a university in Pakistan to predict students' academic achievement at the end of the programme [4]. Compared with overall performance prediction, course performance prediction focusses on predicting the final achievement of a course from e-learning systems or traditional classrooms. The outcome of course performance prediction can help educators providing more knowledge and skills training for students with stronger learning ability and offering more personalized help to students with poor performance. Pressure of students can be reduced as well [5].

Furthermore, the output of course performance prediction can be used as a part and evidence of overall performance prediction.

The main research problem of course performance prediction is the selection of features and algorithms, i.e., specifying related influencing factors as features for specific learning situation and building optimal models by employing proper algorithms for performance prediction. Features of course performance prediction are classified into two categories: learning ability and course involvement. Learning ability includes GPA, grades of prerequisite courses and assignment scores. Course involvement is the participation in course activities, such as the inquiry count. The most commonly used algorithms of course performance prediction are multiple regression, support vector machine, neural network, random forest and logistic regression. For example, Huang et al. took GPA, grades of prerequisite courses and assignment scores as features to predict the performance of students in an engineering dynamics course and found that, compared with multiple regression and neural network, support vector machine was the optimal algorithm for predicting individual students` performance [6]. Features used in the He et al. study included weekly video watching progress, weekly assignment progress and weekly assignment scores. Logistic regression and transfer learning algorithms based on logistic regression were employed and the results showed that LR-SIM, one of the transfer learning algorithms, had better accuracy and stability [7]. Ahadi et al. studied the course performance prediction of 210 students enrolled in a Java course from the University of Helsinki. With major, GPA, programming experience and assignment scores selected as inputs, Bayesian network, J48 decision tree and random forest were used to detect high-performing and low-performing students. It was found that random forest was the optimal algorithm of which accuracy was 90% [8].

According to the research above, since each performance prediction has its own proper features and algorithms, the study of course performance prediction requires feature selection based on the characteristics of the course and the comparative experiment between lots of algorithms, aiming at building an optimal prediction model. Feature selection from both subjective aspect and objective aspect is required for course performance prediction of basic courses of universities. The perception, e.g. individual interest, is one of the subjective features. Objective features include learning ability and course involvement. Research above focus on selection of objective features, e.g. GPA, grades of prerequisite courses and progress of online learning, and lack the subjective features of students. In addition to objective features, subjective features are required in course performance prediction for basic courses of universities to fully reflect the academic achievement of students.

Based on the analysis above and characteristics of basic courses of universities, a novel course performance prediction method is proposed. The performance of students is divided into 5 levels. GPA, grades of prerequisite courses, assignment scores and the inquiry count are selected as objective features and individual interest is specified as the subjective feature. A comparative experiment was implemented to find the best classifier among support vector machine (with Gaussian kernel or polynomial kernel), BP (backpropagation) neural network, random forest and logistic regression. These data mining algorithms are commonly used in course performance prediction and proved to be accurate. It was found that both the accuracy and AP (Average Precision) of support vector machine (SVM) with Gaussian kernel were the highest and reached 99%. The optimal SVM with Gaussian kernel (Gaussian SVM) is selected to build the course performance prediction model for basic courses of universities, which is a new method for course performance prediction.

2. Feature selection and data collection

The data used in this study includes features and the final performance of 54 students enrolled in a basic course Object Oriented Programming (C++) from a university in Beijing. All of the students were observed without direct intervention. The final performance of students is divided into five levels as shown in Table 1. Aiming at predicting the performance level with high accuracy, features from both subjective aspect and objective aspect are required. Individual interest is selected as the subjective feature and GPA, grades of prerequisite courses including Data Structure and Data

Structure Practice, and assignment scores are specified as the objective features. Selection and analysis of each feature is discussed in this section.

Table 1. Levels of performance.

Level of performance	Related grades` interval
1	[0,60)
2	[60,70)
3	[70,80)
4	[80,90)
5	[90,100]

2.1Objective features

2.1.1Grades of prerequisite courses, GPA and assignment scores. Grades of prerequisite courses, GPA and assignment scores are the reflection of students` learning ability. GPA and grades of prerequisite courses were collected from the office of academic affairs. Assignment scores were offered by the course instructor. Scoring of prerequisite courses were centesimal system while GPA and assignment scores were based on a five-point scale. Grades of prerequisite courses represent students` understanding of the prerequisite knowledge and the more recent the prerequisite courses are, the better they can reflect students` acquisition of prerequisite knowledge [4]. Therefore, grades of two recent prerequisite courses, Data Structure and Data Structure Practice, were selected as the inputs. All of the students enrolled in the Object Oriented Programming (C++) course attended Data Structure and Data Structure Practice on the last semester. The two prerequisite courses have a significant influence on students` programming skills and comprehension of concepts in C++. GPA denotes the learning ability of students before and impacts the study of the new course. Assignment scores concretely show the learning of the current course. As depicted in Table 2, the assignment scores of the Object Oriented Programming (C++) course are the marks of 12 labs.

Table 2. Assignment scores.

Week	The count of 5 points	The count of 4 points	The count of 3 points	The count of 2 points	The count of 1 point
1	4	18	20	12	0
2	3	9	31	11	0
3	6	16	22	5	5
4	7	22	14	5	6
5	3	17	19	11	4
6	2	9	30	8	5
7	4	6	29	6	9
8	7	9	29	3	6
9	13	18	11	7	5
10	5	10	22	12	5
11	1	4	29	7	13
12	0	2	32	9	11

2.1.2The inquiry count. The inquiry count reflects the course involvement [3]. The source of inquiries is related to the educational system of the course. In e-learning systems, the inquiry is done by adding a post in the e-learning platform such as Moodle, while for traditional classrooms, the inquiry is based on face-to-face communication and social network service, e.g. WeChat or Facebook. Since most of the students` inquiries in the Object Oriented Programming (C++) course were from Moodle, WeChat and the comments in lab reports, the inquiry count of each student in WeChat, Moodle and lab reports was collected weekly. Each student had 12 records because there were 12 weeks in the course. The weekly inquiry count of different source is shown in Table 3. WeChat inquiries were collected from the chat history of students, the assistant and the course instructor. Moodle inquiries denoted questions from posts in the Discuss Activity of Moodle. Lab report inquiries were from the "discussion and inquiries" section of lab reports.

Table 3. The inquiry count.

Week	WeChat inquiry	Moodle inquiry	Lab report inquiry	Total
1	0	0	19	19
2	0	8	8	16
3	11	14	13	38
4	5	11	3	19
5	5	0	8	13
6	2	3	6	11
7	2	0	2	4
8	1	1	7	9
9	7	0	5	12
10	0	0	8	8
11	1	0	3	4
12	0	0	4	4

2.2Subjective feature: individual interest

As the student`s perception of a course, interest is classified into two types: individual interest and situational interest. Individual interest is the learner`s interest of the course before attending it [9]. Compared with the situational interest, the impact of individual interest to students` performance, knowledge acquisition and values is stronger and longer [10]. Hence, the student`s individual interest is selected as the subjective feature. The initial interest (individual interest) questionnaire [11] designed by Harackiewicz et al. was modified for assessing the individual interest of students enrolled in the Object Oriented Programming (C++) course. The modified questionnaire was based on a five-point scale, as denoted in Appendix A. Students finished filling out the questionnaire at the first week of the course and 54 effective questionnaires were collected. The average of a student`s point for each item in the questionnaire was specified as the individual interest of the student. The statistics of the 54 questionnaires is shown in Table 4 in which values were accurate to two decimal places.

Table 4. Statistics of individual interest.

Item	Max	Min	Average	Standard deviation	Variance
Value	5.00	1.14	3.90	0.87	0.76

3. Model

The course performance prediction model proposed is formulated in equation (1):

$$f(\boldsymbol{X}^{(i)}) = \Phi(\boldsymbol{X}^{(i)}, \boldsymbol{y}^{(i)}) \tag{1}$$

where the vector $\boldsymbol{X}^{(i)} = [X_1^{(i)}, X_2^{(i)}, X_3^{(i)}, X_4^{(i)}, X_5^{(i)}, X_6^{(i)}]^T$ denotes the inputs of the sample i as listed in Table 5. Φ is the classfier of the model based on the algorithm used. Since classification algorithms used in the model are adapted to binary classification, multi-class classification of the model can be decomposed into binary classification by the OvR method. The vector $\boldsymbol{y}^{(i)} = [y_1^{(i)}, y_2^{(i)}, y_3^{(i)}, y_4^{(i)}, y_5^{(i)}]^T$ is the actual class of the sample i, where $y_k^{(i)} \in \{-1,1\} \mid k \in \{1,2,3,4,5\}$. $f(X^{(i)})$ is the prediction result of the model for the sample *i*. The value of $f(X^{(i)})$ is $[f_1(X^{(i)}), f_2(X^{(i)}), f_3(X^{(i)}), f_4(X^{(i)}), f_5(X^{(i)})]^T$ where $f_k(X^{(i)}) \in \{-1,1\}$ and $k \in \{1,2,3,4,5\}$.

Table 5. Inputs of the model.

Feature	Variable
GPA	X_1
grade of Data Structure	X_2
grade of Data Structure Practice	X_3
individual interest	X_4
assignment scores	X_5
inquiry count	X_6

Algorithms including support vector machine, BP neural network, random forest and logistic regression were selected as the classifier of the model. These algorithms are commonly used and proved to be accurate for course performance prediction. Computing the class with decision functions by training the classifier, and tuning the hyper parameters are required by classification algorithms. The training process and hyper parameter of each algorithm are described in this section.

3.1SVM

The core idea of the SVM algorithm is that input vectors are non-linearly mapped to a very high dimension feature space by kernel functions and a decision hyperplane is constructed to separate the samples of different classes in the feature space [12]. An optimization problem solved by SVM is shown in equation (2) :

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2} \boldsymbol{w}^T \boldsymbol{w} + C\sum_{i=1}^{m} \xi_i \\ &\text{s.t. } \boldsymbol{y}^{(i)}(\boldsymbol{w}^T \phi(\boldsymbol{X}^{(i)}) + b) > 1 - \xi_i \\ &\xi_i > 0 \end{aligned} \tag{2}$$

where m is the number of samples and $\boldsymbol{w}$ is the weight vector. ξ is the slack variable and C is the penalty factor. SVM is divided by the kernel function used, e.g. the Gaussian kernel and the polynomial kernel. The decision function of the polynomial SVM is formulated in equation (3).

$$f_k(X) = sign(\sum_{i=1}^{m} \alpha^{(i)} y^{(i)} [X * X^{(i)} + b]^d) \tag{3}$$

The decision function of the Gaussian SVM is shown in equation (4):

$$f_k(X) = sign(\sum_{i=1}^{m} \alpha^{(i)} y^{(i)} \exp(-\gamma \| X - X^{(i)} \|^2)) \tag{4}$$

where *sign* is the signum function and m is the number of samples. $\alpha^{(i)}$ is the Lagrange multiplier and $k\left(k \in \{1,2,3,4,5\}\right)$ is the class of the decision function. In the polynomial kernel, d is the degree of the polynomial and b is the bias. In the Gaussian kernel, γ is the parameter for constructing the high dimension features. With decision function, the model can decide the class of a sample by its feature vector $X^{(i)}$. C, d, b and γ are the hyper parameters to be tuned in the SVM algorithm.

3.2Logistic regression

Logistic regression is one of the classical classification algorithms and can cope with binary classification by sigmoid function. The loss function of logistic regression is defined in equation (5)

$$L(\boldsymbol{w}) = \sum_{i=1}^{m} \log(1 + \exp(-\boldsymbol{y}^{(i)} \boldsymbol{w}^T \boldsymbol{X}^{(i)})) + C \| \boldsymbol{w} \|^2 \tag{5}$$

where m is the number of samples. $\boldsymbol{w}$ is the weight vector and C is the penalty factor. The tuning of C to minimize $L(\boldsymbol{w})$ and improve the generalization is required by logistic regression for an optimal classification model.

3.3BP neural network

Different from logistic regression, the concept layer is introduced in BP neural network, and the hidden layer, a group of non-linear computing units, is added between the input layer and the output layer. By back propagation algorithm, BP neural network alters the weight and threshold continuously to minimize the loss. The BP neural network model constructed in this study was a three-layer MLP (Multi-Layer Perceptron) [13] model. The number of neurons in the input layer and the output layer were fixed. The number of neurons is tuned to optimize the model, i.e., the number of neurons n_{hidden} is the hyper parameter of the BP neural network.

3.4Random forest

Random forest is a kind of ensemble learning method [14]. With the bagging strategy, random forest consists of lots of tree classifiers generated in parallel and the final decision is made based on the results of these tree classifiers [15]. The unit of random forest is a decision tree. Each unit randomly draw a subset of the training sample with replacement. The subset has the same sample number with the training sample but a lower feature number. A vote is casted among all of the units to attain the result of the “forest”, i.e., an overall classification result. The number of units, i.e., decision trees, and the max number of the subset features have an important influence on the accuracy of the algorithm. Therefore, the number of decision trees n_{tree} and the max number of subset features $max_{features}$ are hyper parameters of random forest to be tuned.

4. Result and discussion

A comparative experiment of five algorithms with specified features was conducted in the macOS High Sierra operating system. All of the algorithms discussed in section 3 were implemented in Python. 648 samples were used as the input with 70% for training set and 30% for test set. The classification model is built by training from the training set and tested by the test set. Accuracy is one of the indicators for assessing the effect of the algorithm, which is formulated in equation (6)

$$p(f(\boldsymbol{X}^{(i)}), \boldsymbol{y}^{(i)}) = \frac{1}{m} \sum_{i=1}^{m} 1(f(\boldsymbol{X}^{(i)}) = \boldsymbol{y}^{(i)}) \tag{6}$$

where $1(x)$ is the indicator function.

Before training, it is necessary to normalize the features to avoid the superiority of a feature for its very large value, because classification algorithms like SVM are sensitive to the difference of features` values. Mean normalization was used to scale the value of features to the interval [0,1] for attaining a more reliable model.

During the training process, combined with the 10-fold cross validation, the Grid Search method was used to tune the hyper parameters of the five algorithms. The main idea of Grid Search is choosing a set of values for each hyper parameter, assembling every possible pair of values and finding out the proper pair for the best effect of the model [16]. The tuning result of each algorithm is listed in Table 6.

Table 6. Hyper parameters of 5 models.

Parameter	Gaussian SVM		Polynomial SVM		BP neural network	Random forest		Logistic regression
	C	γ	C	d	n_{hidden}	n_{tree}	$max_{features}$	C

Value	9	0.52	91	3	45	3	2	20

With the hyper parameters altered, the mean accuracy of 5 course performance levels was calculated for each algorithm. As depicted in Figure 1, the mean accuracy of the Gaussian SVM, i.e., SVM(rbf), was the highest, followed with the BP neural network (BPNN) and random forest (RF). The mean accuracy of the three algorithms above reached 95%, while the mean accuracy of the polynomial SVM, i.e., SVM(poly), was low, followed with logistic regression.

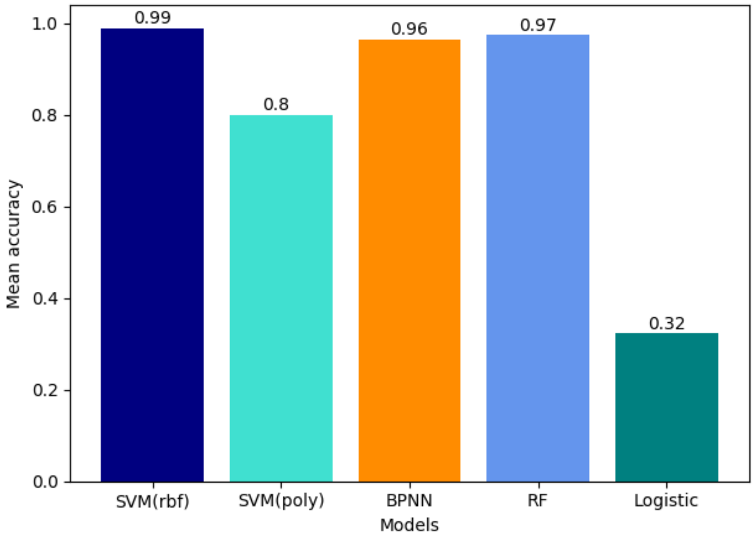

Figure 1. Mean accuracy of the five models.

In addition to accuracy, the ROC (Receiver Operating Characteristic) curve and the PR (Precision-Recall) curve are commonly used to assess the effect of classification models and the PR curve is more proper for skewed samples [17]. In this study, the number of samples with performance level five were much smaller than the number of other samples, i.e., the samples were skewed. Therefore, the PR curve was selected to assess the effect of the model. The definition of precision and recall are shown as follows [18]:

$$precision=\frac{TP}{TP+FP} \tag{7}$$

$$recall=\frac{TP}{TP+FN} \tag{8}$$

where TP, FN and FP denote the relation between the prediction result and the actual result. The relation is called the confusion matrix, as shown in Table 7.

Table 7. Confusion matrix.

Actual→ / Predict↓	true	false
true	TP	FN
false	FN	TN

In the PR curve, the horizontal axis denotes recall and the vertical axis is precision. The area under the PR curve is AP (Average Precision), as defined as follows:

$$AP=\int_0^1 p(r)dr \tag{9}$$

where p is precision and r is recall. The closer to the right top the PR curve is, i.e., the bigger AP is, the better the prediction model is. PR curves of the five algorithms employed in this study are represented in Figure 2.

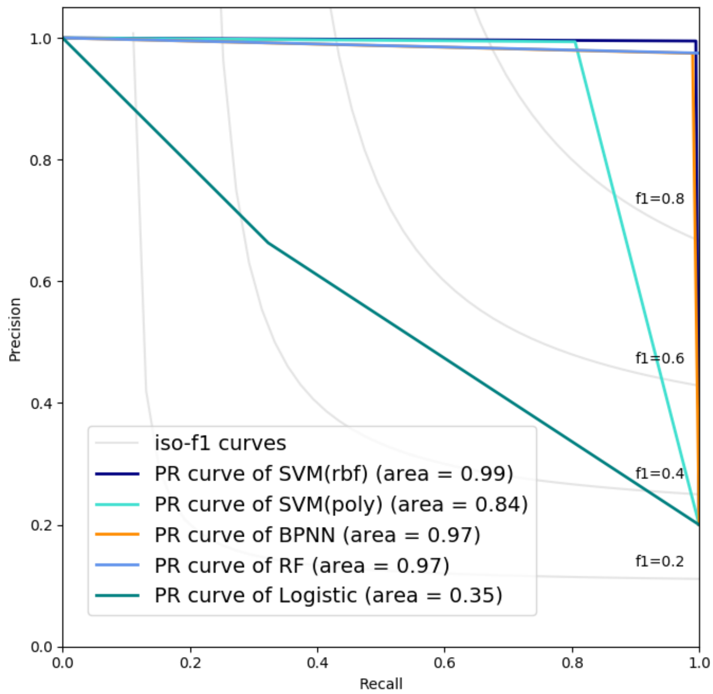

Figure 2. PR curves of five models.

The PR curve of the Gaussian SVM model is the closet to the right top with the highest AP, reaching 0.99. It is implied that with skewed samples, the Gaussian SVM model can predict the course performance better than other four models. It is proved that SVM is adapted to small-sample learning [19]. Therefore, the course performance prediction model with the Gaussian SVM algorithm as the classifier can better predict the course performance level of basic courses of universities.

5. Conclusion

Based on characteristics of basic courses of universities, a course performance prediction model is built by feature selection from both subjective and objective aspects. Individual interest is selected as the subjective feature and GPA, grades of prerequisite courses, assignment scores and the inquiry count are specified as the objective features. The course performance is divided into 5 levels for better assessment of students` performance. After analysing 648 samples from 54 students enrolled in a basic course Object Oriented Programming (C++) of a university in Beijing, this study compared the accuracy and the PR curve of 5 algorithms (i.e., the Gaussian SVM, the polynomial SVM, BP neural network, random forest and logistic regression) commonly used in course performance prediction. Aiming at achieving high accuracy and solving the small-sample and skewing problem caused by the long period of basic courses of universities, the Gaussian SVM algorithm, for its highest accuracy, largest AP and best small-sample adaptability, is chosen as the classifier of the course performance prediction model for basic courses of universities. The model can predict students` final performance from both subjective and objective aspects, which can be a reference for teachers when altering educational strategies. Successive studies include: (a) selection of the knowledge units` acquisition as the objective feature for fully reflection of students` learning; (b) collection of more samples for better generalization; (c) improving of current algorithms to adapt to small-sample learning for higher accuracy.

Acknowledgments

This research is supported by China National Nature Science Foundation (61672361) and Beijing Advanced Innovation Center for Imaging Technology (BAICIT-2016004).

Appendix

A. The questionnaire for assessing the individual interest

Question	Value (disagree)				Value (agree)
I've always been fascinated by Object Oriented Programming (C++).	1	2	3	4	5
I chose to take Object Oriented Programming (C++) because I'm really interested in the topic.	1	2	3	4	5
I'm really excited about taking this class.	1	2	3	4	5
I'm really looking forward to learning more about Object Oriented Programming (C++).	1	2	3	4	5
I think the Object Oriented Programming (C++) is a very important subject.	1	2	3	4	5
I think the field of Object Oriented Programming (C++) is an important discipline.	1	2	3	4	5
I think what we will study in Object Oriented Programming (C++) will be important for me to know.	1	2	3	4	5
I think what we will study in Object Oriented Programming (C++) will be worthwhile to know.	1	2	3	4	5

References

[1] Romero, C., Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert systems with applications, 33(1), 135-146.
[2] Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. Expert systems with applications, 41(4), 1432-1462.
[3] Superby, J. F., Meskens, N., Vandamme, J. P. (2006). Determination of factors influencing the achievement of the first-year university students using data mining methods. In 8th international conference on intelligent tutoring systems, Educational Data Mining Workshop. Jhongli. 32: 234
[4] Asif, R., Merceron, A., Ali, S. A., Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. Computers & Education, 113, 177-194.
[5] Thai-Nghe, N., Horváth, T., Schmidt-Thieme, L. (2011, July). Factorization Models for Forecasting Student Performance. In EDM. Eindhoven. pp. 11-20.
[6] Huang, S., Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. Computers & Education, 61, 133-145.
[7] He, J., Bailey, J., Rubinstein, B. I., Zhang, R. (2015, January). Identifying At-Risk Students in Massive Open Online Courses. In AAAI. Austin. pp. 1749-1755.
[8] Ahadi, A., Lister, R., Haapala, H., Vihavainen, A. (2015, July). Exploring machine learning methods to automatically identify students in need of assistance. In Proceedings of the eleventh annual International Conference on International Computing Education Research. Omaha. pp. 121-130.
[9] Renninger, K. A. (1992). Individual interest and development: Implications for theory and practice. The role of interest in learning and development, 26(3-4), 361-395.
[10] Hidi, S. (1990). Interest and its contribution as a mental resource for learning. Review of Educational research, 60(4), 549-571.
[11] Harackiewicz, J. M., Durik, A. M., Barron, K. E., Linnenbrink-Garcia, L., Tauer, J. M. (2008). The role of achievement goals in the development of interest: Reciprocal relations between

achievement goals, interest, and performance. Journal of educational psychology, 100(1), 105.
[12] Cortes, C., Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.
[13] Gallinari, P., Thiria, S., Badran, F., Fogelman-Soulie, F. (1991). On the relations between discriminant analysis and multilayer perceptrons. neural networks, 4(3), 349-360.
[14] Dietterich, T. G. (2000, June). Ensemble methods in machine learning. In International workshop on multiple classifier systems. Berlin. pp. 1-15
[15] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
[16] Bergstra, J., Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb), 281-305.
[17] Davis, J., Goadrich, M. (2006, June). The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning. Pittsburgh. pp. 233-240.
[18] Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.
[19] Xuegong, Z. (2000). Introduction to statistical learning theory and support vector machines. Acta Automatica Sinica, 26(1), 32-42.