Journal of Engineering and Applied Sciences 13 (6): 1347-1353, 2018
ISSN: 1816-949X

Modeling Student Knowledge Retention Using Deep Learning and Random Forests

[1]N. Sharada, [1]M. Shashi and [2]Xiaolu Xiong
[1]Department of CS and SE, College of Engineering, Andhra University, Visakhapatnam, India
[2]Worcestor Polytechnic Inititute, 100 Institute Road, Worcestor, MA 01609, USA

Abstract: With accurate and dependable student knowledge retention models both over practice and under practice can be avoided in adaptive educational systems. The knowledge retention being a hidden parameter and hence can only be estimated from student responses to the retention tests conducted over a period of time. While the relationships among the independent variables describing the student experiences is nonlinear and complex, the existing personalized scheduling systems attempted to model them using traditional methods such as linear regression and basic statistical techniques. In this research, the application of the most advanced computational techniques such as deep learning and random forest are investigated on the dataset from Personalized Adaptive Scheduling System (PASS) module of ASSISTment, web based mathematics tutor. Experiments demonstrate that deep learning technique for student knowledge retention modeling significantly outperformed the baseline GLM Model with an R^2 value of 0.542. In addition, these techniques are further explored to schedule personalized retention tests after the students initial skill mastery. For this regression problem, the random forest regression technique, indicated a prediction improvement with an R^2 of 0.912 than a baseline linear regression model with an R^2 of 0.417.

Key words: Adaptive educational systems, educational data mining, intelligent tutoring system, knowledge retention, traditional method, PASS

INTRODUCTION

Human brains have a tendency to forget newly acquired knowledge and skills if conscious efforts are not being made to retain it. Forgetting happens regardless of the material being taught or the age or the background of the learner. However, learning is considered robust if the acquired knowledge or skill is retained for longer periods of time. Long term retention of knowledge is important for a variety of reasons. The retrievable information forms a basis for successful transfer of knowledge from one concept to another and for the development of many other cognitive processes. For more than a century, psychologists have identified that temporal expanding retrieval practices lead to more robust and durable learning. The technique involves attempting to retrieve the knowledge immediately after it has been mastered and then gradually increasing the spacing interval between successive retrieval practices. Although, teachers recognize the importance of spaced retrieval practice, the demands of reviewing and re-teaching all the old concepts compete with the imperative to regularly introduce new concepts. Henceforth, a system of Personalized Adaptive Scheduling System (PASS) (Xiaolu and Beck, 2013) was developed to review the initially mastered skills in Assistments, a nonprofit, web based tutoring system. PASS assigns the student a reassessment test on mastered skill at expanding intervals first within 7 days, then 14 days after the prior test, then 28 days later and finally, 56 days later. The first level retention test conducted within 7 days after skill mastery is personalized, based on the student's skill mastery speed. Although, PASS helps the student to review the knowledge after a time period, however, it does not take into account the individuals memory strength for each skill.

The underlying rational of any kind of instruction is the assumption that the skills once mastered are accurately recalled in some other context at some time in future. Nevertheless, most of the research in educational data mining is focused on predicting students short term performance. Corbett and Anderson's (1994) knowledge tracing model has been widely used to model students current knowledge and learning. The model assumes that each skill has two learning parameters, prior and learn and two performance parameters, slip and guess. There is also another parameter, forget which is assumed to be zero in standard knowledge tracing model. Some researchers have carried out work on predicting the students long term retention performance. To account for forgetfulness, knowledge tracing model is extended (Qiu *et al*., 2010) to predict the student's responses after the elapse of a day

Corresponding Author: N. Sharada, Department of CS and SE, College of Engineering, Andhra University, Visakhapatnam, India

or more, since, the previous response. An experimentation was performed to examine the effects of practice and spacing (Pavlik and Anderson, 2005) on retention and it has been observed that wide spacing of practice provides increasing benefit as the practice accumulates. In another research, mastery learning concept (Yutao and Beck, 2012a, b) is extended by considering the long term learning effect on the student performance. The research also identified that the long term retention knowledge is likely to vary by skill and by student. The existing students models (Yutao and Beck, 2012a, b; Xiong *et al*., 2015) focused on identifying the factors responsible for student retention performance using basic statistical techniques such as General Linear Model (GLM). These methods are ideal for identifying simple linear relationships in the data. The proposed personalized retention test scheduler aims to predict the optimum time interval for scheduling the retention test, so that, a particular student would most benefit from recalling. Recently, there is a major advancement in training densely connected, feed forward neural networks with many hidden layers. The deep networks, thus, resulting, learn a hierarchy of nonlinear features which can capture complex patterns in data. These advances triggered researchers interest in developing student models based on deep learning techniques and random forests for estimating knowledge retention.

Problem and approach: Student knowledge retention modeling is crucial for an intelligent tutoring system in order to adapt to the learning needs and the knowledge levels of the individual students. With accurate and dependable student knowledge models both over-practice and under practice can be avoided. The challenge inherent in estimating student performance of a skill is that knowledge retention is a hidden variable and must be inferred from student behavior patterns. Deep neural networks and random forests are effective in estimating the upper boundary of explained variance when there is a complex nonlinear relationship among covariates. The previous study using regression models (Sharada *et al*., 2016) revealed that student individual differences in retaining learned concepts, skill nature and class overall performance should also be considered for estimating retention performance. Hence in this research, the investigation is further extended by building student models using advanced data mining techniques like deep learning and random forests. Accordingly the primary goal of the current study is to identify the best classification algorithm for estimating student long term retention of an already mastered skill.

Learning is considered robust if the acquired skill is retained for long period of time. The cognitive scientists (Carpenter and DeLosh, 2005) has demonstrated two key principles for robust learning. The first principle is the spacing effect which take into account repetition of the study material over spaced intervals of time. The second is the testing effect which involves practicing of retrieving the information from memory. These two principles are combined in order to maximize the benefits known as spaced retrieval practice and is implemented in PASS module of assistments system. PASS reviews the retention of an already mastered skill by scheduling periodic retention tests based on their individual speed of mastering a skill. However, mastery speed alone is insufficient to predict retention. The current study formed a hypothesis that memory retention jointly depends on features like skill latent difficulty, students latent retention ability and the overall study history of the class. Henceforth, the second objective of this study is to develop regression models using the aforementioned features to automatically schedule personalized retention tests. The main objectives of this study are:

- Find the best regression algorithm for customizing student retention test schedules
- Find the best classification algorithm for predicting the long term retention performance

MATERIALS AND METHODS

The web based tutoring system: ASSISTment is a web based mathematics cognitive tutoring system developed for middle school students to provide instruction while simultaneously assessing them thereby avoiding lost instruction time during assessment. The system assist the students while practicing the questions related to a skill until mastery threshold is attained which is defined as three consecutive correct answers without using feedback. PASS is a new module of assistment platform which automatically reassess the students on a schedule with expanding spacing intervals ranging typically from 1-4 levels of around 7, 14, 30 and 60 days where the student has to demonstrate mastery of the skill. The retention test of level 1 is customized based on the students mastery speed which is defined as the number of practice problems required to obtain three consecutive correct answers. When the students mastery speed is three then the retention test is scheduled on 7th day, however, if the student require seven or more opportunities to attain skill mastery, then the retention test is planned the very next day. If the student

demonstrated retention on level 1 test, the skill was reassessed 14 days later, followed by 30 and 60 days. When a student didn't demonstrate retention, feedback was given and practice continued until the student exhibit mastery of the skill. Expanding spaced interval between repeated tests ought to promote student learning and optimal long term retention.

About PASS dataset: The data used in this analysis is taken from the level 1 retention test performance of PASS module in assistments platform. The system is mostly used for urban school districts of the Northern United States for 4th through 10th grade mathematics. The data is collected from school year 2014-2015 which comprises of learning experiences of about 12, 238 unique students while solving problems related to 154 mathematics skills within assistments. In total there are 1, 85, 904 data records, the description of the dataset is in Table 1. Each row of the data set is recorded after a student mastered a skill and the logged information include the identity of the student, the class to which he/she belong to, the identity of the teacher, the skill identity, the mastery speed, number of days after which retention test is conducted, the difficulty of the question that was asked in the test, the result of the retention test in terms of 0 and 1.

About preprocessing: From the fundamental features such as students mastery speed and retention performance, the average skill mastery s peed and average skill retention specific to each skill were computed. This helps to analyze the impact of skill nature information on the retention performance. To study the effectiveness of teaching learning process experienced by the students in the class, the average class mastery speed and the class retention performance were computed. Similarly, the individual differences in student learning and retention are captured by means of average student mastery speed and average student retention performance. When there are outliers in the dataset they could skew the results, hence, transformation is applied to minimize their influence. The maximum number of problems allowed to be answered for a skill on a single day in skill builder problem sets of assistments is 10. Accordingly the mastery speeds <10 in the data set are transformed to 10 to fairly account student retention performance. Nevertheless the impact of this transformation is less significant, since, 95% of the mastery speeds are <10 in the dataset. Two sets of attributes are used for predicting long term retention performance and estimating customized retention test schedules. The data set has an attribute correct which can take values 0 and 1 indicating whether the student has correctly answered the retention test. The successful completion of retention test confirm the estimation made for delay days, implying the length of duration the student can retain the mastered skill. Hence, for estimating the level 1 retention test schedule the rows corresponding to successful completion of retention test are considered. The PASS dataset with 1, 85, 904 records is randomly split into training set and test set in the ratio of 75:25 for further analyses.

Table 1: A students learning experiences description table

Attributes	Description
Student_id	Student identity
Student_class_id	Class identity
Mastery_speed	Number of practice problems required to obtain three consecutive correct answers
Problem_id	Problem identity
Delay_days	Number of days after which the retention test is scheduled
Teacher_id	Teacher identity
Skill_id	Skill identification
Correct	Retention test performance
Skill_ms	Average mastery speed of each skill
Student_ms	Average mastery speed of each student
Class_ms	Average mastery speed of each school
Skill_rt	Average retention performance of each skill
Std_rt	Average retention performance of each student
Class_rt	Average retention performance of each class

Classification and regression algorithms: In the PASS dataset, the student is considered to retain a skill if he/she correctly answers the retest conducted after an estimated delay days. To accomplish the task of predicting long term retention performance, binary classification techniques are applied on the training data and these are further cross-validated by test data. The dependent variable in these models is the performance of the student in the retention test, the incorrect response are treated as '0' and correct response as '1'. To address the second research issue regression models are developed which automatically schedule personalized review tests based on the students individual knowledge levels and learning patterns. The predictor variable is delay days which defines when to assign level 1 retention test after initial skill mastery. The algorithms chosen for the purpose are deep learning and random forests. Before proceeding for model building the conceptual aspects of these algorithms are discussed.

Deep learning: The deep neural networks used for supervised predictive modeling of student retention knowledge are based on multi-layered feed forward networks with multiple layers of interconnected neurons. The computational models (Cun *et al*., 2015) with multiple processing layers transform representation of data at one level to data representations at much higher abstract level. With the composition of many such transformations, complex patterns can be learned. Deep learning

improved many aspects of modern society from speech recognition, visual object detection to recommendations on e-Commerce websites. The multiple layers of neurons present in the feed forward neural network constitute the depth of the network and the number of neurons in each layer represents the width of the network. The weights linking the neurons and biases from other neurons in addition to the width and depth of the network determine the output of the entire network.

To model the student retention performance, the deep learning algorithm is trained with three hidden layers and each layer comprise of 100 neurons. The hyperbolic tangent function is used to transmit the input information, through hidden layers until it reaches the output nodes. The hyperbolic tangent (tanh) function is a rescaled version of the sigmoid function whose output ranges from 1+1. This allows the algorithm to converge faster and is given by:

$$f(a) = \tanh(a) = \frac{e^{\alpha}-e^{-\alpha}}{e^{\alpha}+e^{-\alpha}}$$

Learning occurs when the weights are adapted to minimize the error on the labeled training data. More specifically, for each training example t, the objective is to minimize a loss function, L (W, B|t). Here, W is the collection $\{W_i\}1{:}N\text{-}1$ where W_i denotes the Weight matrix connecting layers i and i+1 for a network of N layers. Similarly, B is the collection $\{b_i\}$ 1:N-1 where $b_i$denotes the column vector of biases for layer i+1. For each training sample 't' the weights are adapted such as to minimize the cross entropy loss function which is given by:

$$L(W, B|t) = \sum_{-Z\in l} (O_z^t * P_z^t + (1-O_z^t) * \ln(1-P_z^t))$$

Where:

W = The weight matrix
B = The column vector of biases
z = The output units
l = The output layer
P and O = The predicted output and actual output, respectively

Binomial distribution function is used along with cross entropy or log-loss for the response variables in the classification. To minimize the log-loss, parallel versions of Stochastic Gradient Descent (SGD) with shared memory is frequently used. However, this creates an extra overhead associated with locking. Hence, in this research a computationally competent lock-free approach is used to parallelize the SGD.

Random forest: A random forest model (Breiman, 2001) with 500 unpruned classification trees are built with a random subset of candidate features. Each tree of the random forest is constructed with a random replacement bootstrap sample from the data. To reduce the correlations among trees, the square root of the number of variables in the data set are randomly sampled for classification and the number of predictors divided by three are considered for regression. The random forest model predicts the category with the average number of votes across all trees for regression and majority of trees for classification.

RESULTS AND DISCUSSION

In this study, it has been illustrated how deep learning and random forests classification techniques can be used in place of traditional approaches such as logistic regression. The primary goal of this research was to improve the performance and efficiency of long term knowledge modeling that will allow the ASSISTments tutor to monitor the student knowledge retention and tailor the retention tests to the students needs. Experiments have been performed using PASS data set obtained from level 1 retention test to evaluate the performance and usefulness of the deep feed forward neural network and random forest models. The approach is developed and analyzed using H_2O an open source library for the R environment. The deep feed forward model is trained with an input layer, three hidden layers (each layer with 100 neurons) and an output layer. The distribution function of response variable is set to binomial and the cross entropy loss function is chosen for model estimation. Random Forest (RF), a powerful ensemble classification algorithm is built with 300 trees as base classifiers. The diversity of the trees in the RF increases by making them grow from different trained datasets created through bagging or boot strap aggregating. The training dataset is analyzed on both of the models. The model goodness is evaluated in terms of log loss, RMSE, R2, AUC, Gini and mean per class error. Each evaluation metric account for different aspects of the model and the data, hence a combination of metrics are ideal for comparing models and assessing the quality of predictions. The log loss or logarithmic loss is a robust classification metric which measures the performance of a model and is defined as:

$$L(a, p) = -a\log(p)-(1-a)\log(1-p)$$

Average log loss for N instances is:

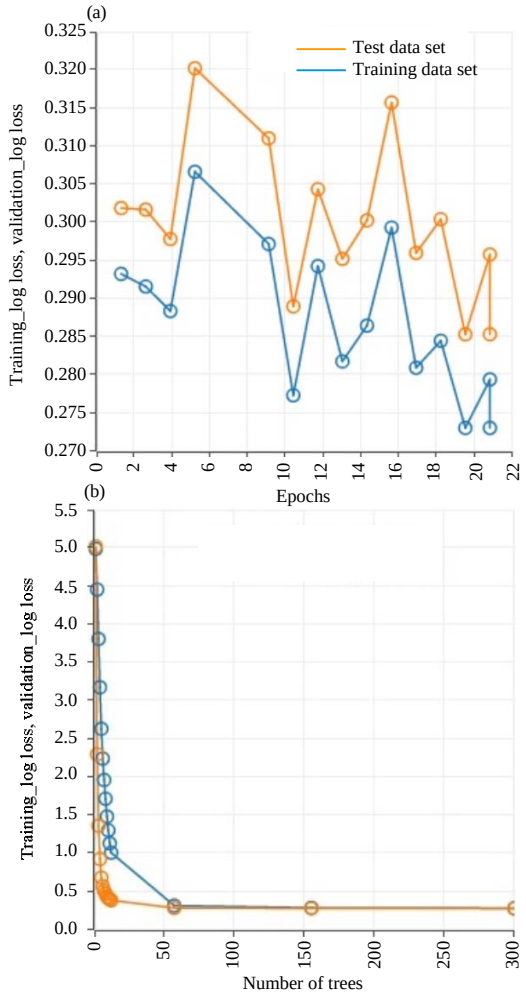

Fig. 1: a) Log loss metric for deep learning and b) Log loss metric for random forests

$$\frac{1}{N}\sum_{i=1}^{N} L(a_i, p_i)$$

Where:
a = The actual value
p = The predicted value

Log loss quantifies the accuracy of the classifier by penalizing wrong classifications. Minimization of log loss leads to maximization of classifier accuracy. In Fig. 1a and b, the log loss error is analyzed with respect to the number of epochs in feed forward neural network and number of trees in random forest, respectively. The accuracy of feed forward neural network with a log loss error of 0.271 is much better than random forest model.

The area under the receiver operating characteristics curve, known as AUC is a standard method for evaluating the predictive accuracy of classifier systems with higher

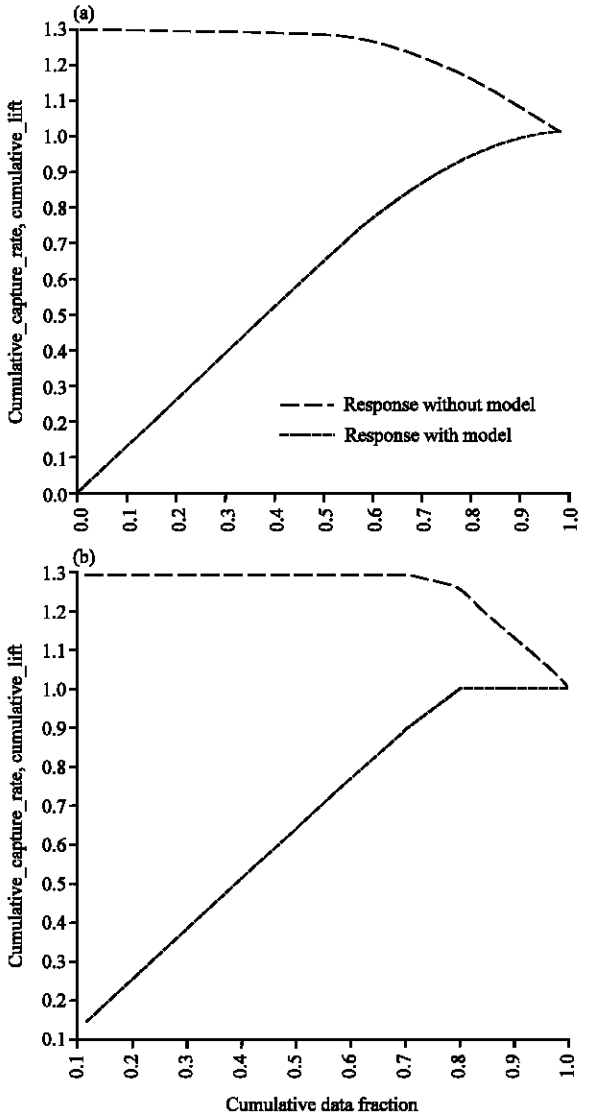

Fig. 2: a) Cumulative gains/Lift charts for DL and b) Cumulative gains/lift charts for RF

scores indicating higher accuracies. The predictor variable for student retention performance is a binary variable with two possible values of 1 and 0 indicating correct and incorrect responses, respectively. Hence, the performance of a binary classifier is measured by plotting true positive rate against the false rate. Neural network achieved significantly higher prediction accuracy than RF with AUC of 0.933 (RF with AUC of 0.922), indicating deep learning is best student knowledge retention model.

Lift is a measure of the effectiveness of the predictive model which is calculated as the ratio of the results obtained with and without the predictive model. Cumulative gains and lift charts (Fig. 2a and b) are visual aids for measuring model performance. The greater the

Table 2: Evaluation metrics for each of the classification models

Classification metrics	Logistic regression	Deep learning	Random forest
MSE	0.081	0.072	0.0916
RMSE	0.432	0.268	0.3820
R^2	0.312	0.542	0.4710
Logloss	0.341	0.271	0.2850
AUC	0.832	0.933	0.9225
Gini	0.932	0.832	0.8450
Mean per class error	0.346	0.195	0.2350

area between the cumulative gain curve and the baseline is the better the model. DL has achieved significantly greater model performance than RF in terms of gains/lifts chart measure as well.

The model performances are also assessed using the error metrics such as MSE, RMSE and coefficient of determination (R^2). The Root mean square error is a measure of average deviation of the predicted values from the observed values whereas R^2 measure the variability in the depended variable explained by the regression model.

Each models effectiveness in predicting student retention is summarized in Table 2. The evaluation metrics such as log loss error function which measures models classification accuracy and AUC which indicates models ability to produce the target value close to predicted value demonstrate that DL performed better than other models. Hence, it is clear that deep learning out performs RF and linear regression for long term retention performance of student learners with evident improvements in all measurements.

Regression analysis: With the encouraging results on classification problem, an attempt is made to predict personalized first level retention test schedules using deep learning and random forest regression models. The data is randomly divided into training set (with 70% of data) for fitting the models and remaining 30% of data for testing the model. To assess the predictive performances, 22 epochs were applied on feed forward neural network with 3 hidden layers (each 100 neurons). This model performance is compared with random forest regression model constructed with 300 trees in Fig. 3a.

The model performance is plotted in terms of mean residual deviance vs. epochs in DL (Fig. 3b) and mean residual deviance vs. number of trees in RF. The lower the value the better is the model. These findings suggest that RF Model has comparatively better student retention test schedule prediction accuracy.

The primary goal of this research was to improve the performance and efficiency of long term knowledge modeling that will allow the assistments tutor to monitor the student knowledge retention and tailor the retention tests to the students needs. The model performance

Table 3: Evaluation metrics for each of the regression models

Regression metrics	Deep learning	Random forest	Linear regression
MSE	0.608	0.339	2.27
RMSE	0.78	0.582	1.508
R^2	0.849	0.912	0.417
Mean residual deviance	0.608	0.339	2.275
Mean absolute error	0.316	0.191	0.974
Root mean squared log error	0.167	0.122	-

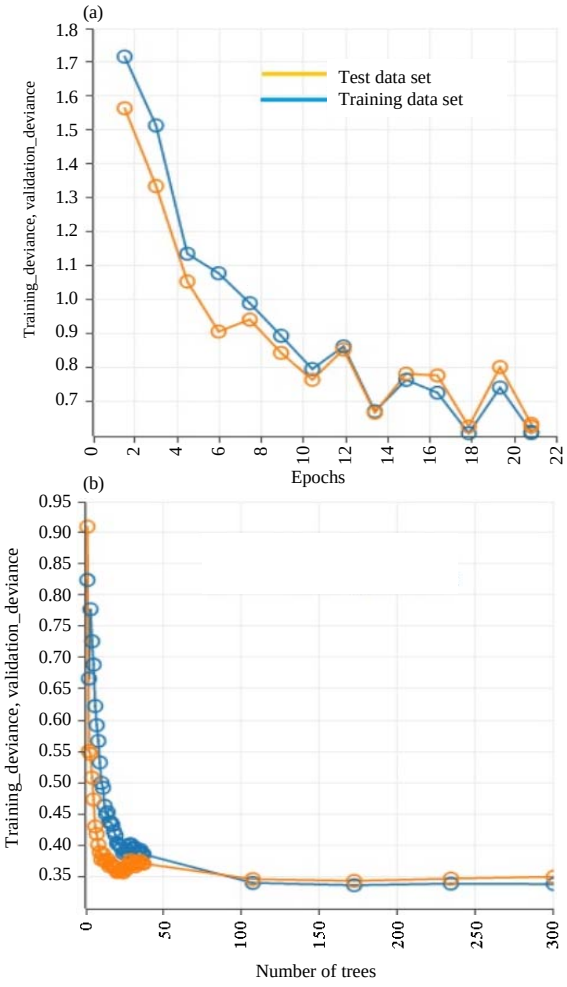

Fig. 3: a) Deviance vs. epochs in DL and b) Deviance vs. trees in RF

metrics in terms of error metrics and coefficient of determination (R^2) are summarized in Table 3. The ensemble random forest regression model achieved better results. This suggest that PASS should incorporate RF technique to enhance the precision of predicting delay days.

CONCLUSION

This study the application of most advanced computational techniques such as deep learning and

random forests are investigated for modeling and predicting student knowledge retention. The objective of this study is to determine whether these models are better approaches for predicting the student retention performance. The performance of deep learning classification technique has been found to be promising for handling the uncertainty in the retention performance. There are many educational objectives for classifying student retention performance, some of them are to identify students with lower motivation and initiate remedial action to lower drop outs, to identify potential student groups with similar characteristics, etc. In the next section a new approach is investigated for scheduling the personalized retention tests that can be used to complement standard methods such as linear regression. The results indicate Random Forest Model performed significantly more accurate in modeling and predicting student retention test schedules than the deep learning regression methods trained on the same data.

More issues could be addressed to refine the accuracy of the student knowledge retention model such as how to explicitly represent time in deep learning to take into consideration the changes in the student's knowledge due to forgetting. And in the long, precise estimation of student knowledge retention will enable more accurate assignment of retention tests to students, thereby optimize the amount of reviews on each skill and may even enable different types of remediation for different types of retention levels.

REFERENCES

Breiman, L., 2001. Random forests. Mach. Learn. J., 45: 5-32.

Carpenter, S.K. and E.L. DeLosh, 2005. Application of the testing and spacing effects to name learning. Appl. Cognit. Psychol., 19: 619-636.

Corbett, A.T. and J.R. Anderson, 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. User Model. User Adapted Interact., 4: 253-278.

Cun, L.Y., Y. Bengio and G. Hinton, 2015. Deep learning. Nat., 521: 436-444.

Pavlik, P.I. and J.R. Anderson, 2005. Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. Cognit. Sci., 29: 559-586.

Qiu, Y., Y. Qi, H. Lu, Z. Pardos and N. Heffernan, 2010. Does time matter? Modeling the effect of time with bayesian knowledge tracing. Proceedings of the International Conference on Educational Data Mining, November 6, 2010, EDM Publisher, Singapore, pp: 1-10.

Sharada, N., M. Shashi and X. Xiong, 2016. Enhanced retention performance modeling for intelligent tutoring system. Intl. J. Comput. Appl., 151: 1-4.

Xiaolu, X., S. Li and J.E. Beck, 2013. Will you get it right next week: Predict delayed performance in enhanced ITS mastery cycle. Proceedings of the 26th International Conference on Florida Artificial Intelligence Research Society FLAIRS, June 10-14, 2013, AAAI, Menlo Park, California, pp: 533-537.

Xiong, X., Y. Wang and J.B. Beck, 2015. Improving students long-term retention performance: A study on personalized retention schedules. Proceedings of the 5th International Conference on Learning Analytics and Knowledge, March 16-20, 2015, ACM, Poughkeepsie, New York, ISBN: 978-1-4503-3417-4, pp: 325-329.

Yutao, W. and J. Beck, 2012a. Incorporating factors influencing knowledge retention into a student model. Proceedings of the International Conference on Educational Data Mining, June 19-21, 2012, Panorama Hotel, Dubai, UAE., pp: 1-4.

Yutao, W. and J.E. Beck, 2012b. Using student modeling to estimate student knowledge retention. Proceedings of the International Conference on Educational Data Mining Society, Jun 19-21, 2012, Springer, Chania, Greece, pp: 444-453.