A Study on the Application of Machine Learning to Learning Analysis

Kim, Yeonhee
Chungbuk National University
kimyh@hoseo.edu

Byun, Hoseung
Chungbuk National University
Hobyun@cbnu.ac.kr

Abstract: Learning analysis first appeared in 2000. An international consortium has been formed and a worldwide standard process is being created as research has been conducted in various fields. In addition, studies are underway to apply machine learning to the analysis of learning data, focusing on some researchers. In this study, we collected learning data from 960 students and studied them using machine learning algorithms and verified their accuracy using test data. The accuracy of the data was verified through 169 test data as a pattern of machine learning obtained from 690 people (80%). For accuracy, the error rate of actual credits and patterns was calculated. As a result, It was noted that with more than 400,000 times of study, the margin of error was collected at 8.4 %.

Keywords: Learning Analytics, Machine Learning, Tensorflow, Linear Regression

INTRODUCTION

The fourth industrial revolution is changing all industries and all areas of human life. In the field of education, the form of learning is changing by OCW and MOOC. In the field of professor design, ICT is applied to learning processes or various teaching strategies that enhance learning effectiveness for learners. Learning analytics is the measurement, collection, analysis and reporting of data related to learners and their situations in order to understand and optimize learning and learning environments.

It is possible to personalize learning by collecting various data and providing profiles such as behavior and personality as well as student performance. In some cases, schools analyze data and provide predictive analytics for customized education and learning. Recently, some studies have been conducted to apply machine learning to the data analysis process. Therefore, we will examine the predictability through machine learning to collect learning data that is deal with learning analytics.

RESEARCH DESIGN & METHODS

In this paper, machine learning was applied to analyze learning data as shown in Figure 1 below. For research, we collected data of 960 students In college. Learning data for creating predictive models should be extracted directly from the data accumulated in the Learning Management System, but data collection has been replaced by surveys because it is difficult to access data related to personal information.

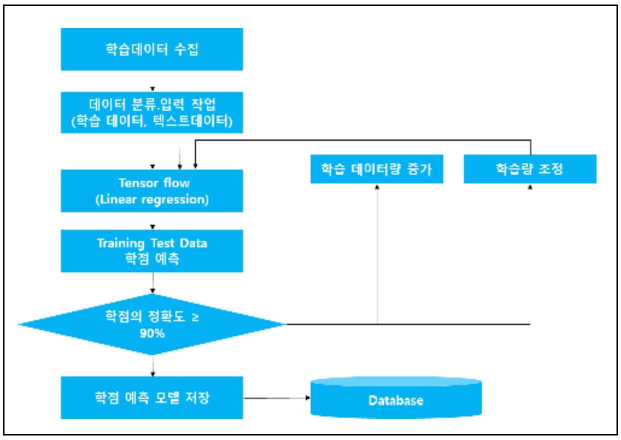
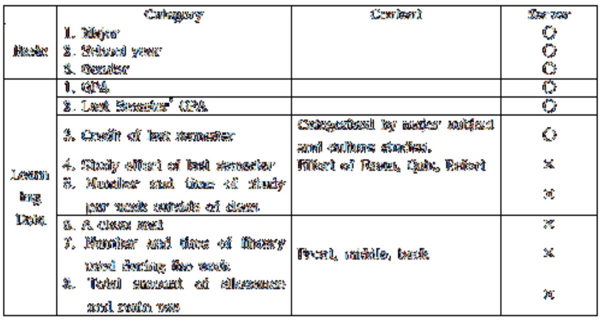


Figure 1. Flowchart of Machine Learning

There are a lot of data related to students ' learning such as Private learning time, School attendance status, Spare time in addition to the credit you get from the server. Therefore, the data related to learning from this survey are as follows.

Table 1. Survey Item



	Category	Content	Server
Basic	1. Major 2. School year 3. Gender		○ ○ ○
Learning Data	1. GPA		○
	2. Last Semester' GPA		○
	3. Credit of last semester 4. Study effort of last semester 5. Number and time of study per week outside of class	Categorized by major subject and culture studies. Effort of Exam, Quiz, Report	○ × ×
	6. A class seat 7. Number and time of library used during the week 8. Total amount of allowance and main use	Front, middle, back	× × ×

The number of people surveyed was 960 but 101 were excluded. The Data for 859 people were used and analyzed. 80 % were used as training data and 20 % as test data. I used machine learning to predict the credit of last semester and checked its accuracy. In statistics, linear regression is modeling that models relationships between dependent variables y and one or more independent variables x. Used program for machine learning is TENSORFLOW. It is made by Google to open the source in 2015. In a linear regression, it is the cost function that minimizes the difference between the hypothesis and the actual value. The cost function is called the least squares method and is used as follows.

$$H_{(x)} = W * x + b$$

$$Cost(W, b) = \frac{1}{m}\sum_{i=1}^{m}(H(x^{(i)}) - y^{(i)})^2$$

Here, w is weight, x is an independent variable, H(x) is hypothesis, y is the result of the smallest W and b values, which makes smaller the difference between the actual and the hypothesis value. this makes accurate predictions. Based on the learning data, After creating machine learning algorithm and function F, when an unlearned input vector $\hat{x}$ is entered, the output $\hat{Y}$ is inferred automatically.

RESEARNCH METHOD

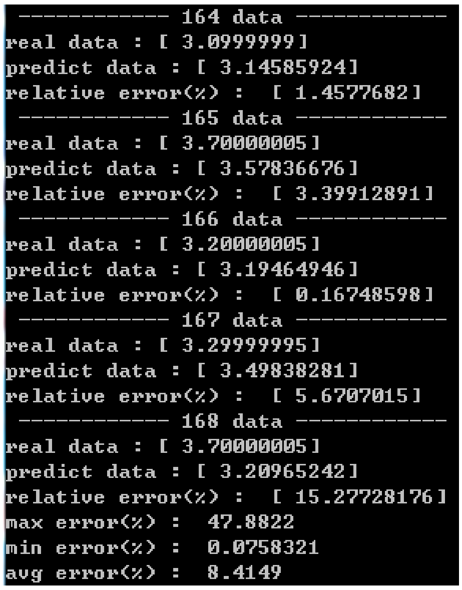
In this study, machine learning is used to analyze learning data. Learning is conducted using an artificial intelligence library called Python and Tensorflow. Learning data is imported into training data and test data as shown below.

```
xy = np.genfromtxt('training_data.csv', delimiter=',',

dtype=np.float32)

test = np.genfromtxt('test_data.csv', delimiter=',',

dtype=np.float32)

-----

W = tf.Variable(tf.random_normal([9, 1]), name='weight')
b = tf.Variable(tf.random_normal([1]), name='bias')

hypothesis = tf.matmul(X, W) + b
cost = tf.reduce_mean(tf.square(hypothesis - Y))
optimizer =
tf.train.GradientDescentOptimizer(learning_rate=1e-5)
train = optimizer.minimize(cost)

sess = tf.Session()
sess.run(tf.global_variables_initializer())

for step in range(1000001):
cost_val, hy_val, _ = sess.run([cost, hypothesis, train],
feed_dict={X: x_data, Y: y_data})
```

The cost is a function of calculating the difference between the estimate value of the whole sample and the actual value Y. The Gradient Descent Optimize is an error minimization process in which the weight W and the calculated predicted value are associated with the actual value Y. In a graph drawn with the value of the cost function and W, the slope calculation sets the next W value for which the cost may be smaller. From the collected data, data for 859 people were used for the study data items shown in Table 1 to pass the previous semester's rating to Y and other data items to X values. After setting up the first Hypothesis equation, the process is repeated. As the cycle is repeated, the value of W is reduced by the minimization process and the value of W is gradually changed from any value to an array associated with the actual value. Set an equation for obtaining the prediction function through the hypothesis value, Set an equation that calculates the average error by using the cost function. And the value is optimized using the GDA and the alpha value of 0.0005. As the weight array are learned to better predict, the w value requires the weight array as a learning result, not temporary value.

RESULTS

Learning data in learning analytics should collect all data related to learning activities, including the physical behavior of students. So we conducted a survey to collect data related to learning that did not exist on the school server. For comparative analysis of predictive models, 690 people in 80 % of 859 people were used as training data and 169 people in 20 % of 859 people were used as Test data and then analyzed the patterns.



```
 ------------- 164 data -------------
real data : [ 3.0999999]
predict data : [ 3.14585924]
relative error(%) :  [ 1.4577682]
 ------------- 165 data -------------
real data : [ 3.70000005]
predict data : [ 3.57836676]
relative error(%) :  [ 3.39912891]
 ------------- 166 data -------------
real data : [ 3.20000005]
predict data : [ 3.19464946]
relative error(%) :  [ 0.16748598]
 ------------- 167 data -------------
real data : [ 3.29999995]
predict data : [ 3.49838281]
relative error(%) :  [ 5.6707015]
 ------------- 168 data -------------
real data : [ 3.70000005]
predict data : [ 3.20965242]
relative error(%) :  [ 15.27728176]
max error(%) :  47.8822
min error(%) :  0.0758321
avg error(%) :  8.4149
```

As a result of comparing hypothesis and actual credit, the patterns of the learning number seems are shown according to the number of studies are shown in Figure 2 below.

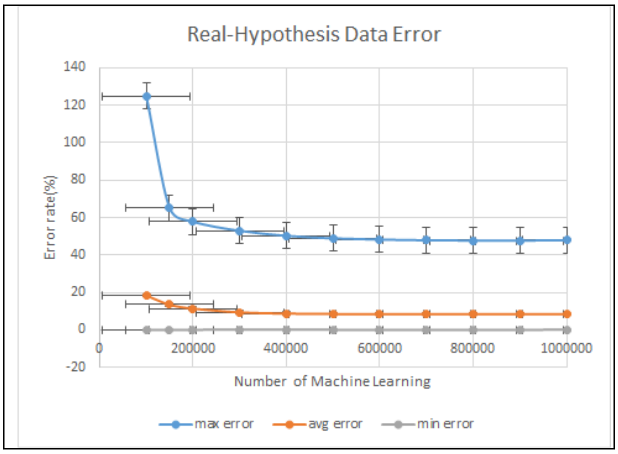


Figure 2. Final results based on learning number of times

As such, we used linear regression algorithms for machine learning to learn and check the results. As a result, it was found that after 400,000 times, the margin of error was collected at about 8.4 %.

Table 2. Real-Hypothesis Value Error of Test Data

Error. Learning Num	Maximum Error	Average Error	Minimum Error
100,000	124.879	18.3669	0.0075255
150,000	65.0897	13.4384	0.0019445
200,000	57.8722	11.3799	0.0178076
400,000	50.3141	8.67916	0.0494712
600,000	48.2326	8.42804	0.0064584
800,000	47.7652	8.43927	0.0402067
1,000,000	47.8822	8.4149	0.0758321

CONCLUSION

After the study analysis, there are many ways to analyze and utilize learning data. In this study, We collect learning data from students, through linear regression of machine learning and then learn and confirm whether they had predicted it or not from the test data. In this study, we studied the learning data through linear regression of machine learning and checked the test data to see if it was predicted. After checking the accuracy to of machine learning was checked by increasing the number of learning, the error rate was collected at about 8.4%. If you remove the higher error rate people of the test data and relearn it, the error rate will be smaller.

Therefore, this study has drawn conclusions that Learning Analytics by machine learning was possible to predict learning result. Although this study did not directly extract learning data from the server and did not collect sufficient learning data. There are many difficulties in getting learning data from the university's learning management system server but it is expected to be possible in the near future. Also we will have to think about how to collect learning data that does not exist on the server.

REFERENCES

H. Benli, “Performance prediction between horizontal and vertical source heat pump systems for greenhouse heating with the use of artificial neural networks,” Heat and Mass Transfer, vol. 52, no.8, pp. 1707-1724, 2016.

K. Han, W. Lee, and K. Sung, “Development of a model to analyze the relationship between smart pig farm environmental data and daily weight increase based on decision tree,” *Journal of Korea Institute of information and communication engineering*, vol. 20, no. 12, pp.2348-2354, Dec. 2016.

S. Shahinfar, D. Page, J. Guenther, V. Cabrera, P. Fricke and K. Weigel, “Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms,”*Journal of dairy science*, vol. 97, no. 2, pp.731-742, Feb. 2014.

M. A. Kashiha, C. Bahr, S. Ott, C. P. Moons, T. A. Niewold, F. Tuyttens and D. Berckmans, “Automatic monitoring of pig locomotion using image analysis,” *Livestock Science*, vol. 159, no. 1, pp.141-148, Jan. 2014.

K. Kim, K. Kim, J. Kim, K. Seol, J. Hong, Y. Jung, J. Park, and Y. Kim, “Changes of serum electrolytes and hematological profiles in Yorkshire at a high ambient temperature” *Journal of Agriculture and Life Science,* vol. 49, no. 1, pp.103-113, Nov. 2014.

Grégoire, G. "Multiple linear regression." European Astronomical Society Publications Series 66 (2014): 45-72.

Govan, Anjela. "Introduction to optimization." North Carolina State University, SAMSI NDHS, Undergraduate workshop. 2006.

Weaver, Jesse, and Paul Tarjan. "Facebook linked data via the graph API." Semantic Web 4.3 (2013): 245-250.

Bakharia, A., & Dawson, S. (2011). SNAPP: A bird’s-eye-view of temporal participant interaction. Proceedings of the First International Conference on Learning Analytics and

Knowledge-LAK'11 pp. 168-173. New York, NY: ACM Press.

S. J. Roberts, R. Cain and M. S. Dawkins, "Prediction of welfare outcomes for broiler chickens using Bayesian regression on continuous optical flow data,“ *Journal of the Royal Society interface*, vol. 9, no. 77, pp.3436-3443, Sep. 2012.

M. S. Lee and Y.C. Choe, "Forecasting Sow's Productivity using the Machine Learning Models," *Journal of Agricultural Extension & Community Development*, vol. 16, no. 4, pp. 939-965, Dec. 2009.