Yang, J. C. et al. (Eds.) (2018). Proceedings of the 26th International Conference on Computers in Education. Philippines: Asia-Pacific Society for Computers in Education

Towards Final Scores Prediction over Clickstream Using Machine Learning Methods

Mohammad Nehal HASNINE[a*], Gokhan AKCAPINAR[a, b], Brendan FLANAGAN[a], Rwitajit MAJUMDAR[a], Kousuke MOURI[c] & Hiroaki OGATA[a]
[a]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
[b] *Department of Computer Education & Instructional Technology, Hacettepe University, Turkey*
[c]*Institute of Engineering, Tokyo University of Agriculture and Technology, Japan*
*hasnine.mohammadnehal.5z@kyoto-u.ac.jp

Abstract: E-books are capable of producing a significant amount of clickstream data that insights students' learning behavior. Clickstream data are often analyzed in learning analytics and educational data mining domains to understand students' synchronous and asynchronous learning processes. The present study analyzed a dataset consisting of university students' clickstream data for predicting their final scores using machine-learning methods. To begin with, the raw data are preprocessed in four steps, namely data aggregation, feature generation, data balancing, and feature selection. After that, utilizing machine learning methods, high performing and low performing students' final scores are predicted. For this, eight machine-learning methods (Neural Network, AdaBoost, Logistic Regression; Naïve Bayes, kNN, Support Vector Machine, Random Forest, and CN2 Rule Induction) are employed and their performances were compared. Result revealed that CN2 Rule Induction algorithm having 88% accuracy outperformed other machine learning methods when best-5 selected features from the dataset were taken into consideration. However, the Multilayer Perceptron based Neural Network performed best having the similar accuracy with CN2 Rule Induction when all features were considered to predict. This paper also focuses on how SMOTE as a data balancing algorithm can be applied to solve data imbalance problem and various scoring methods can be compared to identify the most important feature attributes in clickstream.

Keywords: Clickstream Analysis, e-Book, Educational Data Mining, Final Score Prediction, Learning Analytics

1. Introduction

Concerning education, a dramatic growth in the adaptation of digital teaching material delivery system (synonymously electronic book) has observed over the last two decades. The adaptation of using digital teaching material delivery systems in higher education is growing because these systems are convenient. Researchers use these systems because they are capable of generating a vast amount of interaction data. These interaction data are known as clickstreams (i.e. mouse clicks). However, as educational data mining matures, students' synchronous and asynchronous learning processes can be examined with clickstream data. A major application of prediction in educational data mining is to predict students' educational outcomes (Asif et al., 2017). In this context, students' final performance prediction has gained increased emphasis in higher education (Romero et al., 2010, Xing et al., 2015). An objective of final score prediction in higher education is, instructors can monitor students' progress and identify at-risk or low performing students in order to provide feedback (Xing et al., 2015). Clickstream data are often used for this purpose together with the early prediction of final performance (Akçapınar et al., 2015), detect drop-out (Dekker et al., 2009), early prediction (Chen W et al., 2018), learning behavior analysis (Chen X et al., 2018), provide feedback (Chen X et al., 2018), intervention (Herder et al., 2018) etc. Furthermore, clickstream data can be transformed into knowledge that could help academicians, administrators, and policymakers to analyze it to enhance decision-making (Asif et al., 2017). While the adoption of the digital textbook in higher education is growing, drawbacks of using digital textbooks are also reported. With this regard, previous studies (Bigot & Rouet, 2007) (Dennis, 2011) also reported that, for students, it is much easier to concentrate on the topic when using paper than on a screen, and therefore printed materials helps to remember and understand information more precisely. Although continuous argument remains in favor of and against adopting digital textbooks in higher education, the importance of having student-generated data cannot be overlooked.

The present works intend to shed light on students' final scores prediction utilizing clickstream data that are collected from an e-book system. The aim of this work is to provide answers to the following research questions: (1) How can a rather small clickstream dataset be analyzed and modeled so that it can be used as a measure for students' final score? (2) How to compare entire features –vs- top-ranked features to enhance prediction accuracy? (3) How to improve learning algorithms in the case of a small sample and imbalanced data?

2. Dataset

The dataset is available to download from the workshop webpage under agreeing upon certain terms and conditions. The dataset contains clickstream data from university students (N=53) from the period of 2017-11-22 to 2018-01-29. The dataset is collected from students' engagement in three different books in a digital teaching material delivery system (hereafter e-book system) (Ogata et al., 2015) (Flanagan & Ogata., 2017). In this e-book system, students can highlight difficult and important areas, take memos, search specific contents, jump to next page, return to the previous page, bookmark important pages and so on. Table 1 shows a summary of the students' interactions in each book.

Table 1: Insight of the Dataset

Book	Students-engaged	Interactions-counted
1	20	6161
2	24	7959
3	53	14696
Total		28816

3. Data Analysis and Prediction

Machine learning algorithms are much more efficient and capable of handling complex datasets. However, in machine learning discipline, no free lunch theorem yielded that there is no one algorithm that is best for all problems (Whitley et al., 2005). Thus, the choice of a correct algorithm often remains unclear unless we test out diverse algorithms directly through plain old trial and error. Depending on the nature of the dataset, extensive preprocessing may require to make raw data readable for machine learning algorithms. Hence, the first step to our data analysis was data preprocessing. Data are preprocessed in four steps, namely data aggregation, feature generation, data balancing, and feature selection.

3.1 Preprocessing

3.1.1 Data Aggregation

In order to preprocess the raw data (clickstream), the present study adapted data aggregation method. According to IBM Knowledge Center, data aggregation is a process where raw data is gathered and expressed in a summary form for statistical analysis. The summary of the data can be in the forms of average, sum, count, maximum, and minimum etc. Aggregated data let data scientists gain insights about the particular data source. We aggregated data using SQL (Structured Query Language) commands. In data aggregation, we counted the total number of events for each feature.

3.1.2 Feature Generation

The attributes in the raw data are *userid*, *action*, *operationname*, *markercolor*, *processcode*, *devicecode*, *markerposition*, *markertext*, *operationname*, and *pageno*. Employing data aggregation method, twenty-five feature attributes were generated. Table 2 shows the list of newly generated feature attributes. R programming language was used to generate new features from the raw data. In the analysis, features in (*) are eliminated because there are not enough data. Note that, newly-generated features are co-related with each other because the source of the raw data is clickstream from the same e-Book system. For this kind of data analysis, finding independent variables are difficult.

Table 2: List of Newly-generated Feature Attributes

Feature Attributes				
contentcount	uniqueday	jump	delredmarker*	delmemo*
sessioncount	longeventratio	bookmarkjump*	yellowmarker	bookmark
totaltime	next	searchjump*	delyellowmarker*	delbookmark
totalevent	prev	memojump*	memo	search
uniqueweek	open	redmarker*	changememo*	score*

3.1.3 Data Balancing using SMOTE Algorithm

In higher education, some universities (e.g. Kyoto University) set students' passing score to at least 60 out of 100. We considered this as the baseline for our final score prediction-related analysis. Based on this baseline, we manually created two groups of students, namely Low Performer and High Performer. High performers are those who achieved high scores in the final. In contrary, students who are at-risk to obtain the passing score are denoted as low performers. The necessity of data balancing arose when we observed a huge gap in students final score between high and low performing groups. In data mining discipline, this situation is called data imbalance problem that may lead to a low prediction accuracy for machine learning methods. Therefore, our approach was to balance data fairly among high and low performing groups. For data balancing, we incorporated conventional SMOTE (Synthetic Minority Over-Sampling Technique) algorithm. SMOTE is an approach often adopted by researchers for the construction of classifiers from the imbalanced dataset (Chawla et al., 2002). We created sample data for low performing student group using SMOTE algorithm. Figure 1 displays the data distribution before and after applying the SMOTE algorithm.

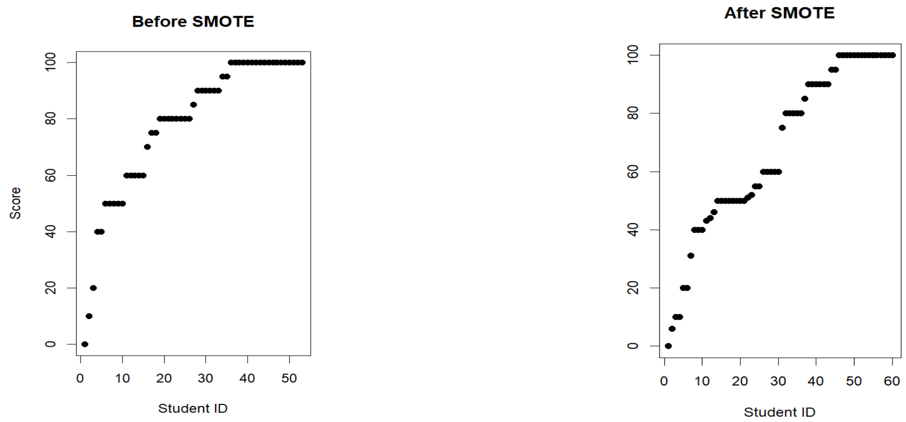

Figure 1: Original Data and SMOTE'd Data

3.1.4 Feature Selection

In machine learning disciplines, feature selection is defined as a process that chooses a minimal subset of features from an original set of features so that the feature space is optimally reduced according to a certain criterion (Novaković, 2016). Diverse feature ranking techniques to discard irrelevant, redundant or unnecessary features from a given feature vector is proposed. However, selecting the most relevant features that fit a model is intractable (Kohavi R., 1997). One of the key objectives of this present study is to analyze and compare different features in order to obtain the best prediction accuracy. Hence, we used the ranking technique utilizing the scoring method. We employed three different scoring methods provided by Orange data mining tool, namely Information Gain (IG), Gain Ratio (GR), Gini Decrease (GD) to the datasets. We have elected five most important features because they are the most important features found in all three scoring methods. Those five most important features were *sessioncount*, *uniqueday*, *open*, *yellowmarker*, and *bookmark.*

3.2 Employing Machine Learning Methods

We have employed eight methods to the processed dataset and compared their performance and prediction accuracies. The eight machine learning methods are, (1) Multilayer Perceptron based Neural Network; (2) AdaBoost; (3) Logistic Regression; (4) Naïve Bayes; (5) kNN; (6) Support Vector Machine; (7) Random Forest; and (8) CN2 Rule Induction. We used 5-fold cross-validation technique for the dataset to compare the performance of different machine learning models. The process of employing machine learning methods is shown in Figure 2. We used Orange[1], a data mining tool for this analysis.

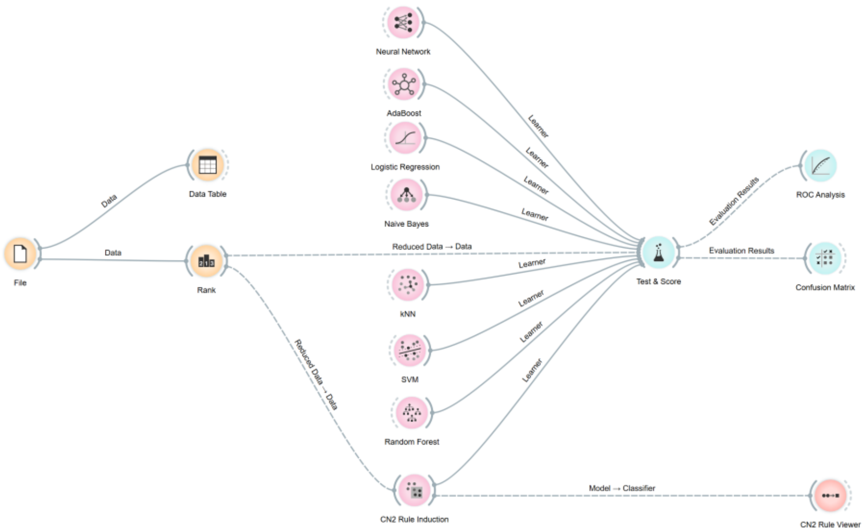

Figure 2: Process of Employing Machine Learning Methods

3.3 Result

Result yield that CN2 Rule Inducer outperformed other methods by having 88% accuracy (AUC=0.93, F1= 0.88, Precision =0.88, Recall=0.88) with 5 most important features. On the contrary, Neural Network performed best when we considered all features of having 88% accuracy (AUC=0.9, F1= 0.88, Precision =0.88, Recall=0.88). Table 3 and Table 4 shows the result of a 5-fold cross validation with 5 important features and all features to predict, respectively. We also analyzed the error rate using a confusion matrix. The confusion matrix in Table 5 demonstrates that CN2 Rule Induce algorithm predicted 10.3% students as High Performer but they are actually Low Performer, and students as Low Performer but they are High Performer, which we addressed as the algorithmic error.

Table 3: Cross Validated Results of Machine Learning Methods with Important Features

Method	AUC	CA	F1	Precision	Recall
CN2rule inducer	**0.939**	**0.883**	**0.883**	**0.884**	**0.883**
Random Forest	0.872	0.817	0.817	0.817	0.817
kNN	0.858	0.767	0.766	0.771	0.767
AdaBoost	0.767	0.767	0.766	0.771	0.767

[1] https://orange.biolab.si/

Neural Network	0.867	0.75	0.749	0.753	0.75
Logistic Regression	0.839	0.733	0.733	0.733	0.733
SVM	0.8	0.7	0.699	0.704	0.7
Naive Bayes	0.842	0.683	0.683	0.685	0.683

Table 4: Cross Validated Results of Machine Learning Methods with All Features

Method	AUC	CA	F1	Precision	Recall
Neural Network	**0.9**	**0.883**	**0.883**	**0.884**	**0.883**
CN2 rule inducer	0.872	0.8	0.8	0.801	0.8
Random Forest	0.856	0.75	0.749	0.753	0.75
Logistic Regression	0.839	0.75	0.75	0.75	0.75
Naive Bayes	0.817	0.683	0.683	0.684	0.683
SVM	0.811	0.667	0.667	0.667	0.667
AdaBoost	0.767	0.767	0.766	0.771	0.767
kNN	0.581	0.433	0.431	0.432	0.433

Table 5: Cross-validated Confusion Matrix for CN2

	Predicted		**Total**
Actual	87.1%	10.3%	30
	12.9%	89.7%	30
Total	31	29	60

Figure 3 displays some rules by set by CN2 Rule Inducer algorithm. The first rule in CN2 Rule Inducer method indicates that- if OPEN is clicked more than 18 times, then 93% probability of all students to get a high score (that is, become high performers). Our analysis also indicated that the overall accuracy of all methods increased for selected features. We yield this conclusion that, feature extraction and feature selection from are recommended. We also suggest that the neural network is a black box algorithm so it is difficult to interpret. In contrary, CN2 is a rule-based algorithm so it is easy to interpret for non-expert users of data mining (such as teachers).

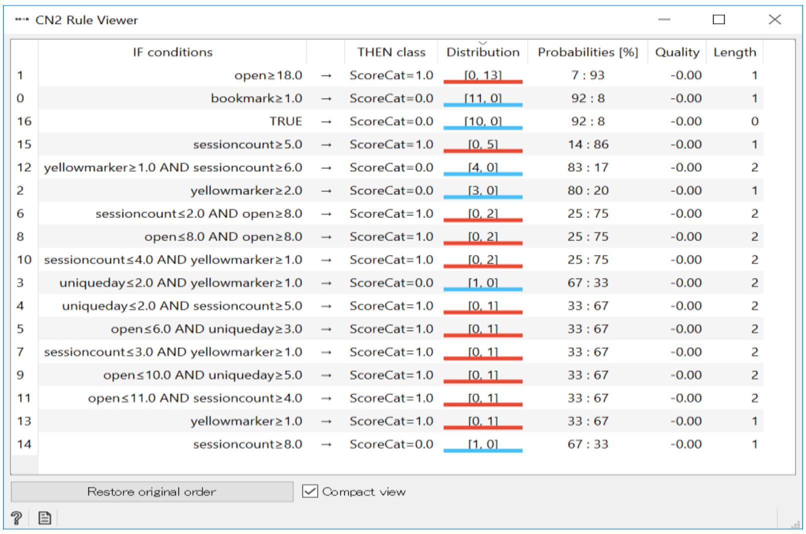

Figure 3: Rules Set by CN2

4. Discussion

Digital e-book systems are used as a tool in university-level education not just for their convenience but also for the vast amount of clickstream data that they produce. Hidden in these clickstreams is valuable information that implies students' behavior of a specific course. Yet from clickstream, identifying common behavior to predict on students' final outcome of a course is a complex and challenging task. In this study, we tried to predict students' final scores based on their clickstream data. In prediction, we employed eight different machine learning methods for high and low performing groups. Eight machine-learning methods (Multilayer Perceptron based Neural Network, AdaBoost, Logistic Regression; Naïve Bayes, kNN, Support Vector Machine, Random Forest, and CN2 Rule Induction) are employed and their performances were compared. Result revealed that CN2 Rule Induction algorithm (88% accuracy) outperformed other methods for best-5 selected features in the dataset. In contrary, Neural Network performed best for all features (88% accuracy). Few important aspects regarding our analysis are: First, we applied SMOTE algorithms to balance the data between high and low performing group. However, one of the issues about SMOTE algorithm that is often discussed researcher is that this algorithm often creates synthetic samples from the minority class instead of creating copies. Second, we assume that approximately 12% error rate arose because some students might have gained high scores in final without relying on the e-Book system. Because student interaction in e-Book is one factor that effects in the final score. There are other factors that may have influence in final scores such as, previous experience with e-Book, learning strategies, learning styles etc. Third, at the beginning of our analysis, we eliminated a certain number of features because there were not sufficient data to represent those features. However, those eliminated features may play significant roles for some students.

Acknowledgements

This research was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (S) Grant Number 16H06304.

References

Akçapınar, G., Altun, A., & Aşkar, P. (2015). Modeling Students' Academic Performance Based on Their Interactions in an Online Learning Environment. *İlköğretim Online*, *14*(3).

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, *113*, 177-194.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Chen, W., Brinton, C. G., Cao, D., Mason-singh, A., Lu, C., & Chiang, M. (2018). Early detection prediction of learning outcomes in online short-courses via learning behaviors. *IEEE Transactions on Learning Technologies*, 1-14.

Chen, X., Breslow, L., & DeBoer, J. (2018). Analyzing productive learning behaviors for students using immediate corrective feedback in a blended learning environment. *Computers & Education*, *117*, 59-74.

Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting Students Drop Out: A Case Study. *International Working Group on Educational Data Mining*, (pp.41-50).

Flanagan, B., & Ogata, H. (2017). Integration of Learning Analytics Research and Production Systems While Protecting Privacy, ICCE 2017 (pp.333-338).

Herder, T., Swiecki, Z., Fougt, S. S., Tamborg, A. L., Allsopp, B. B., Shaffer, D. W., & Misfeldt, M. (2018, March). Supporting teachers' intervention in students' virtual collaboration using a network based model. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 21-25).

Junco, R., & Clem, C. (2015). Predicting course outcomes with digital textbook usage data. *The Internet and Higher Education*, *27*, 54-63.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, *97*(1-2), 273-324.

Lu, L., Dunham, M., & Meng, Y. (2005, August). Mining significant usage patterns from clickstream data. In *International Workshop on Knowledge Discovery on the Web* (pp. 1-17). Springer, Berlin, Heidelberg.

Novaković, J. (2016). Toward optimal feature selection using ranking methods and classification algorithms. *Yugoslav Journal of Operations Research*, *21*(1), 119-135.

Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., & Yamada, M. (2015). E-book-based learning analytics in University education. In *Proceedings of the 23rd International Conference on Computers in Education, ICCE 2015* (pp. 401–406).

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601-618.

Whitley, D., & Watson, J. P. (2005). Complexity theory and the no free lunch theorem. In *Search Methodologies* (pp. 317-339). Springer, Boston, MA.

Xing, W., Guo, R., Petakovic, E., & Goggins, S. (2015). Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory. *Computers in Human Behavior, 47*, 168-181.