International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 5, October 2019, pp. 3569~3575
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i5.pp3569-3575

Proposed classification for elearning data analytics with MOA

Chanintorn Jittawiriyanukoon
Graduate School of Advanced Technology Management, Assumption University, Thailand

Article Info

Article history:

Received Nov 5, 2018
Revised Apr 8, 2019
Accepted Apr 28, 2019

Keywords:

Classification
Data analytics
Elearning
LMS
MOA

ABSTRACT

Elearning education has developed a crucial factor in the educational organization. With the situation of declining student size, elearning has to offer more cross-departmental and multi-disciplinary courses for individual needs to go over "one-size-fits-all" traditional model. Elearning data analytics which has not been professionally classified cannot produce reliable results. Classifications for elearning data help comfort the accuracy of outcomes and reducible pre-processing time. This research proposes a practical model for individual learning and personality. The proposed model based on data from the LMS classifies both the student preferences and personalities. The model helps design future curricula to suit student personalities, which intangibly assists them to be efficient in the study practice. Performance of the proposed classification is evaluated by using MOA software. It outperforms and improves the accuracy of complex elearning datasets. Besides, the results indicate an achievement in the students' study time after applying the association rule model on the elearning.

Corresponding Author:

Chanintorn Jittawiriyanukoon,
Graduate School of Advanced Technology Management,
Assumption University,
Samut Prakan Province, 10540, Thailand.
Email: pct2526@yahoo.com

1. INTRODUCTION

Elearning has developed from the phase of utilizing smart devices to access digital content which is residing on the virtual learning system. With the emergence of cloud-based applications, eLearning is obtaining popularity in digital learning and in academic education. Not to mention, with the internet power available everywhere, data analytics plays a critical role in eLearning environment experiences. Learning Analytics has been about in pedagogy as digital education allows to collect facilitator attention as well as learners' behavioral data, by establishing quizzes and exams. Analytics also plays an ultimate part in learning practices. With the advance of technological innovations, learning is presently dispersed through mobile gadgets. Learners thus employ several social media networks, Tweeter, WhatsApp, emails, WeChat and etc. Collected data is an unstructured format, but still, insights remain. In eLearning environment, data analytics focuses on making use of insights to leverage learning outcomes by using useful tools such as machine learning, regression-based predictive analysis, and similarity mapping, it is likely to design a learning system that satisfies to individual learner needs. After integrating outcomes from data analytics into eLearning environment it results more in-depth and ubiquitous assessment of learners. Knowledge repositories over the website of chat, wikis, blogs and video stream shared by online communities are gradually attracting the main resources for current students. It is significant to account learners' interactions with social knowledge repositories through phases of analytics [1].

Adaptive systems can automate digital learning content through analytics by varying responses to input parameters based upon historical context and collected data from different sources. The efficiency of this system is evaluated by the capacity to improve outcomes regarding these alterations. For example,

systems are designed to collect data, analyze and make decisions to improve learners' experience. The learning experience starts with mouse clicks, navigation, and interaction with online resources. With analytics, who are active learners can be monitored and how they interact with Learning Management Systems (LMS) can be viewed. In the old days, analytics has been driven by the needs of the educational institute in order to support learning and teaching. However, there has been an excess of analytical approaches from different grounds, including data mining, business intelligence, web analytics and predictive modeling. Due to multi-disciplinary learning activities and emergence of "Big Data Curation" [2], there are various prospects to reflect problematic questions and provide solutions to deal with demands of elearning. For example, fully interacting with learners is deliberated as one of the important communication channels to enable learning experience. Mostly, by analyzing data about learners' activities and picturing results, tips on improvement and real-time response about learners can be given.

Data analytics classifies how activities of learners are [3] and how they interact with elearning contents. Having information on such interactions helps facilitators create an exact comprehension of learners' requirements. By allowing facilitators to observe and compare success across various methods of instruction, the impact and quality in the teaching of their courses are raised. The use of big data to boost learning models is growing. A great number of educational institutions are implementing big data analytics in elearning to reform teaching activities. There has been a transformation to a learning activity where data analyst helps provide an effective learning outcome and support decision-making process for both learners and facilitators. Currently, universities are aware of the ecosystem emerging around the data analysis. In this regard, perfect strategies and tools should be familiarized, the right know-how is practically applicable and analyzing how learning can be enhanced is needed. Absolutely, integration of analytical procedures into elearning has unlocked a new arena for innovation [4] which indeed can provide learners and facilitators with the speedy response on their learning processes. This possibility is the source for reforming teaching structures. In reality, it is essential to classify the role of data analytics and its impact on traditional and future learning models.

This research centers on classifying elearning data using MOA simulation [5]. Both pre-processing (classification) and post-processing (association) are executed for collected data. Firstly, the performance has been highlighted to develop pre-processing capacity in terms of classification accuracy and precision. Secondly, an association-based method has been executed, and these experimental outcomes are used to leverage future curricula in order to suit learner personalities and indirectly to aid them to be much more efficient in the study practice. Lastly, conclusion and future works are discussed.

2. E-LEARNING DATASET

In elearning environments, a dataset about learners' engagement can be retrieved. There are needs to cater online courses for more flexible and interactive interfaces to motivate both facilitators and students in a virtual class. The qualified elearning environment provides a real-time interactive response to both facilitators and students during the learning process. To comprehend students' engagement at several steps of the elearning experience can help cater more interfaces that further information learned by students and help initialize the learning experience and reduce dropout rates. This paper uses a synthetic dataset which contains students' engagement during elearning experience.

The synthetic dataset contains 1,800 attributes and 7,000 instances about e-learning environment, explained with labels for student id, engagement, registration, evaluation and background levels. The dataset composes of detailed records of students who currently enroll. A synthetic dataset with 16 attributes and 480 instances also has been employed by [6] for the mining process. Especially, how to apply elearning data analytics to help disabled students has been presented by [7]. It is likely possible that the enrolled subjects may tell students' behavior to suit their objectives of the learning. Carmona et al. [8] demonstrate elearning student preferences in order to make decision for matching some features of the LMS with the student's elearning methodology. Ghatasheh [9] presents an innovative design of an eLearning organization, where user activities and machine learning are taken into consideration. The paper focuses on the evaluation of online students' behavior. Classification algorithms such as features extraction, map analysis, and normalization are introduced to forecast the knowledge level of online students. However, the paper does not evaluate the performance of pre-processing approaches in terms of classification accuracy (CA) and precision. Farrus and Costa-juss´a [10] present an automatic assessment of elearning assignments by emphasizing on its applicability and significance. The authors also point out the favorable results of an automatic evaluation tool based upon Semantic Analysis approach, which has been inspected in a particular elearning environment. Employing an automatic scoring system for an evaluation reflects the online students a prompt access to their achievement at any time, without the assistance of a facilitator. The automatic system as such provides more adaptable for students to timetable and to fine-tune their vulnerable.

3. DATA CLASSIFICATIONS

This section illustrates diverse performance considerations of the classification mechanisms in the elearning dataset. The main purpose is to evaluate classification algorithms, which are taken into consideration in the pre-processing phase. The assessment metrics for the classification algorithms are classification accuracy (CA), precision (PR) and area under the receiver operating characteristic curve (AUROC). The reliable and well known SMO, RF and Zero-R classifiers are investigated then these outcomes are compared to what the proposed algorithm can result. These four algorithms are described as follows.

3.1. Sequential minimal optimization (SMO)

The algorithm is a machine learning approach which trains a support vector classification by replacing all missingness and changing nominal attributes into binary values. The algorithm normalizes all attributes as well as contributes that the coefficient in the outcome depends on the dataset normalization, not the primary content. This approach produces ultimate estimation results, but its accuracy reckons on a set of input parameters. The detail of the algorithm can be found in [11].

3.2. Random forest (RF)

The classification algorithm presented by [12] trains dataset with more than two classes by building a forest of a random decision-tree structure. The algorithm employs randomization on the input parameter as well. The individual tree is established from a root sample of the dataset. As constructing each tree, a subset of attributes is randomly outlined from which superb attribute for the fragment has opted. The difference between the traditional decision tree algorithm and RF is that in the RF, the final decision depends on the popular vote from each tree established in the forest. Moreover, RF can be applicable for data streams, but its performance is not essentially improved [13].

3.3. Zero-R (ZR)

The algorithm [14] calculates the recurrent class, the average value for a numeric class and the mode for a nominal class from the datasets. This classifier develops a model which usually forecasts the majority and average value. In case of more than two majority classes, the classifier selects the class randomly. The algorithm per se is always employed as a basic approach for other models. The simple ZR method builds a frequency table for all possible targets and chooses the recurrent value. It is clearly seen that the algorithm is nothing about the predictions towards the model as ZR does not apply anything.

3.4. Proposed method

The proposed method is an approach combining results from several forecasting models to develop a different model for the sake of higher precision. The method splits the calculation into two levels. The first level, the integration of inputs from each of the individual classifiers after the execution is taken into consideration in order to find out each baseline model whose performance outperforms. The second level, each baseline model whose performance is inferior will be ignored. For this motivation, the proposed algorithm is operational as the baseline models are all diverse. Suppose six people shoot a combination of 200 arrows at a target. Meanwhile, only 50 of those who can shoot at the right target. For the rest, their performance can be discarded. The only task is to find out who did not shoot the arrow on landing target, detach them and keep collecting those 50 outperformers. The algorithm of the proposed method is depicted in Figure 1.

Proposed Method

Require: Dataset matrix $[M]_{ab}$ with a rows and b columns
Ensure:$[M]_{ab}$, C = total number of classifications, T = dimension of [M]
for i = 1 to a do
for j = 1 to b do
  for k = 1 to C do /** First level of prediction **/
    Classifier C_k applying for dataset [M]
  end for
  for n = 1 to T do /** Prediction to maximize regression-based probability**/
  $M_c = \{a'_n, b_n\}$, where $a'_n = c_0 + c_1 a_n + c_2 a_n + \ldots + c_T a_n$
  end for
    Choose new classifier C with M_c /**Second level of prediction **/
    Return C
end for
end for

Figure 1. Proposed method prediction

4. EXPERIMENTAL MODEL

Cloud-based LMS can contribute an extensive scale of prospects, for instance, cost-effectiveness followed by a wide-ranging scalability. The secure-cloud-based LMS issues a number of online services in connection with social media and mobile applications for the learners' experience. In parallel, there are dedicated services to log the learner behavioral activities. The accumulation of the exam performances and the activity records of the individual learner have been logged in a data repository. The repository then is employed in the assessment of the online learners. Veselinova and Ristova [15] have developed advanced technologies and innovation using open-source based Moodle, in the elearning classroom, for the sake of making the language virtual classroom more attractive and creative as well as motivating learners to utilize this pedagogy with their facilitators. Facilitators can decide soft-skill objectives and further construct them in a framework of successive activities for drafting a lesson plan based upon a communicative competence of elearning students as studied in [16].

In this paper, the experimental model as depicted in Figure 2 consists of three sections. The pre-processing based upon classification technique is the first process of extracting an insightful data from the repository. The insight from this stage is warehoused for a further process, a step of association. Finally, the suitability of the offered courses to the learners regarding standard curricula in online educational programs is analyzed. As concluding outcomes from the third step, the dynamic elearning environment and adaptive elearning system must be considered to mimic the learners' behavior of elearning based method.

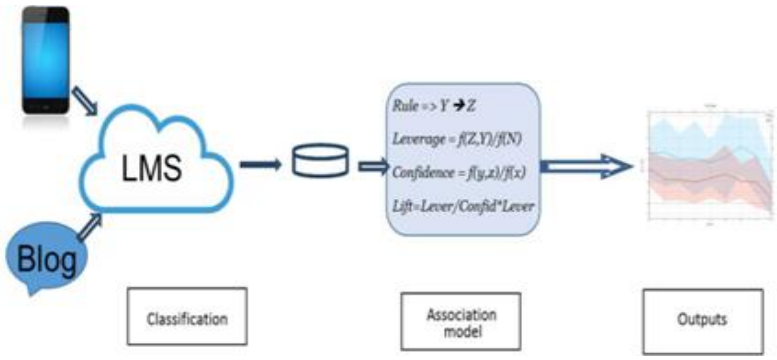

Figure 2. Experimental model

5. RESULTS AND ANALYSIS

The classification models are constructed and evaluated in this paper utilizing the synthetic datasets from an elearning system (so-called LMS). The individual synthetic dataset denotes the learner's model and is mainly categorized into three basic parts which are the examination result attributes, the individual registration records and the background knowledge records. There are about 7,000 instances whose size is about 0.2 GB. In this paper, the MOA simulation is employed for the data analytics. All experiments have been run on an Asus Windows 7 with Intel® Core ™ i7 CPU, 2.2 GHz Processor and 8 GB RAM. The datasets have been opted in different years and with variable size. The experimental model as shown in Figure 2 has been executed in order to collect performance metrics (such as AUROC, precision, and CA) of the proposed algorithm then compare to other three traditional algorithms, namely SMO, RF, and ZR. The pre-processing improvement results from five datasets in the repository are listed in Table 1.

Numerous researches aim at emerging methods for association rule. But, simplification and efficiency are obstructions for the development as such. Normally, pre-processing (classification) and post-processing (association model) are both necessary for curating data, such as extracting and converting the data from any particular format to expected format. In this paper, the first attempt has been completed to propose an effective classification comparatively among the SMO, RF, and ZR approaches. The next attempt involves with association model to achieve valuable information. Association rule model intends to curate insightful, frequent patterns, inter-correlations, or any associations between sets of instances in datasets or other databases from the repository [17]. The main objective of the model is to find out the set of all attributes or instances in which regularly repeat in database transactions or records, and furthermore, to check rules on whether or not instances subset affects the existence of other subsets. Association rule algorithm concerns high-level estimation rules based on the *If-Then-Else* condition; *IF* the condition of the estimation attribute is "True", *THEN* estimate value for the targeted attribute. In addition, the rule is a relation

between two occurrences in the pattern of A→ C, in which A is anticipated while C is a successor. The rule per se demonstrates how many occurrences C has followed after A arises regarding the confidence and leverage values.

Table 1. Results from classification

Dataset1			
Classifier Type	Precision (%)	AUROC	CA (%)
SMO	94.4	98	93
RF	95	98.2	95.2
ZR	42.1	50	65
Proposed	97.4	99.2	97
Dataset2			
Classifier Type	Precision (%)	AUROC	CA (%)
SMO	94.6	98.7	94.1
RF	93.2	97.7	93.1
ZR	43.2	50	62
Proposed	96.9	99	96.8
Dataset3			
Classifier Type	Precision (%)	AUROC	CA (%)
SMO	91.6	98	91
RF	93.1	97.5	93.2
ZR	42.4	50	65.9
Proposed	96.2	98.8	96.9
Dataset4			
Classifier Type	Precision (%)	AUROC	CA (%)
SMO	65	50.7	62.1
RF	84.1	82.9	84.7
ZR	57.6	50	75.9
Proposed	84.5	89.1	85
Dataset5			
Classifier Type	Precision (%)	AUROC	CA (%)
SMO	97	98	97.5
RF	97.2	97.8	97.2
ZR	27	50	52
Proposed	98.2	98.8	98.5

The specific formulas which are employed to calculate the association metrics for a given association rule (A→C) are *support*, *confidence* and *leverage. Support* represents the probability of instances in the dataset that holds an itemset in the given database where *support* (A→C) is defined as *support* (A U C). *Confidence* denotes a conditional probability, for a rule, where *confidence* (A→C) can be described as *support* (A→C)/ *support* (A). *Leverage* is the difference between the expected probability of the independent itemset and the probability of the rule and *leverage* (A→C) can be characterized as *support* (A→C) – (*support* (A)* *support* (C)).

The Apriori algorithm, so-called the *level-wise* algorithm, is applied for the association model (post-processing). The algorithm per se is based upon preceding information of repeated itemset properties. The algorithm employs two states (an entrant and a clipping) in each recursion. It keeps iterating by starting from the first level till the n^{th} level, in which no entrant leftovers after discarding.

In the preliminary simulation in post-processing, to list all association rules, the *confidence* value has been initially set to be 1 %, and the *support* degree is set to be 0%. After the association rule execution, 315 association rules, as well as 72 rules whose *confidence* degree is larger than 60%, have been achieved. Table 2 demonstrates how these 315 rules are distributed. The maximum *confidence* value of the rule descends when the number preceding itemset drops, as the number of preceding itemset is identical to 9, the maximum *confidence* value is about 90%, while the *confidence* value of all association rules is lower than 60%, where the maximum *confidence* is around 43%. The following results designate that a student involves more on social media; he will have higher opportunity for elearning graduation in specified duration of time.

Association rules whose *confidence* value is larger than 60% are rules in which the number of precedent items is 7. Thirty-nine rules explain 51.2% of rules whose *confidence* value is larger than 60%. Table 3 lists the results with graduation factor as the successor. These association rules are in descending order of *confidence* values. The results shown in Table 3 demonstrate that students who have got high English test score, and complete all core courses, will produce a 100% chance of graduation. The following 100% probability of graduating is also 100%, which arises in those students who have got high GPA at high school, complete all core courses, and take IT project option.

Table 2. Association model using Apriori algorithm between social communication factors and graduation: successor = graduation in specified period of time

Number of preceding set	Number repeated itemsets (conf > 60%)	Number repeated itemsets (conf > 5%)	Confidence (%)		
			Average	Min	Max
1	0	6	5.6	3.4	8.4
2	0	17	10.9	6.6	21
3	0	25	19.7	8.9	33.5
4	0	39	35.1	19.3	48.1
5	0	96	42.9	15.7	60.2
6	14	45	67.2	44.3	78.8
7	39	38	46.8	26.4	80.3
8	10	27	74.0	70.5	85.7
9	6	8	80.6	76.4	90.2
10	2	2	N/A	N/A	N/A

Table 3. Association model among factors: successor = graduation

Sequence	Precedence I	Precedence II	Precedence III	Conf	Lift	Lev	Conv
1	High Eng Score	Core Courses		1	1.2	0.07	13.2
2	High School GPA	Core Courses	IT Project	1	2.49	0.24	47.9
3	High Eng Score	Family Business		1	2.62	0.06	13.0
4	Social Media	Working	Physical Ed	1	2.59	0.23	12.5
5	Average School GPA	IT project		0.98	1.8	0.18	18.2

6. CONCLUSION

This elearning data analytics found that the most significant factor in graduation is English test score, followed by core courses completion, and IT project involvement. In our experimental dataset, 3.53% of the students could not finish. Regarding to the first association rule shown in Table 2, with social communication as the precedent item and graduation in specified time as the successor, the confident value is 8.4% – this is about 2.84 times the occurrence rate of the total training dataset. The experiment has revealed that English test is one of the largest factors of graduation, and is particularly co-operating the group of students who can graduate just in time. Some researchers have appealed that social media may help students succeed in elearning education, students from business family cannot be beneficial entirely, but related project in IT leads to insignificance. The results of this research show that in physical education as the successor, there are 39 motivating association rules. The results point out that short of physical exercise could lead to the associated factors of unfulfillment. Prevention and intervention of fail factors such as incompletion of all core courses can help increase chance of graduation greatly. The direct prevention is changing of studying style, such as focusing on core courses, enjoying regular physical education, and involving more IT projects. Future research will take other association algorithms, such as frequent pattern or genetic algorithms, into account.

REFERENCES

[1] K. Zdravkova, "Reinforcing social media based learning, knowledge acquisition and learning evaluation," *Proceedings of 2nd International Conference on Higher Education Advances*, vol. 228, pp. 16-23, 2016.
[2] C. Jittawiriyanukoon, "Evaluation of a multiple regression model for noisy and missing data," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 4, pp. 2220-2229, 2018.
[3] K. Alsaadat, "Mobile learning technologies," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, no. 5, pp. 2833-2837, 2017.
[4] L. P. Kushnir and K. C. Berry, "Inside, outside, upside down: New directions in online teaching and learning," *Proceedings of the International Conference e-Learning*, pp. 133-140, 2014.
[5] A. Bifet, R. Kirkby, G. Holmes and B. Pfahringer, "MOA: Massive Online Analysis," *Journal of Machine Learning Research*, vol. 11, pp. 1601–1604, 2010.
[6] G. E. Yuliastuti, et. al., "Performance analysis of data mining methods for sexually transmitted disease classification," *International Journal of Electrical and Computer Engineering (IJECE)*, vol.8, no.5, pp. 3933- 3939, 2018.
[7] M. Cooper, R. Ferguson and A. Wolff, "What can analytics contribute to accessibility in e-Learning systems and to disabled students' learning?," *Proceedings of the 6th ACM International Conference on Learning Analytics & Knowledge (ICPS)*, 2016, pp. 99–103.
[8] C. Carmona, G. Castillo and E. Millan, "Discovering student preferences in e-Learning," *Proceedings of the International Workshop on Applying Data Mining in e-Learning*, pp. 33-42, 2007.
[9] N. Ghatasheh, "Knowledge level assessment in e-Learning systems using machine learning and user activity analysis," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 4, pp. 107-113, 2015.

[10] M. Farrus and M. R. Costa-juss´a, "Automatic evaluation for e-learning using latent semantic analysis: A use case," *The International Review of Research in Open and Distributed Learning*, vol. 14, no. 1, pp. 239–254, 2013.
[11] X. Shao, K. Wu and B. Liao, "Single directional SMO algorithm for least squares support vector machines," *Computational Intelligence and Neuroscience*, pp. 1-7, 2013.
[12] N. Zimmerman, A. A. Presto, S. P. N. Kumar, J. Gu, A. Hauryliuk, E. S. Robinson, A. L. Robinson and R. Subramanian, "A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring," *Atmospheric Measurement Techniques*, vol. 11, no.1, pp. 291-313, 2018.
[13] H. Abdulsalam, D. B. Skillicorn and P. Martin, "Streaming random forests," *Proceedings of the 11th International Database Engineering and Applications Symposium*, pp. 225–232, 2007.
[14] C. Nasa and Suman, "Evaluation of different classification techniques for WEB data," International Journal of Computer Applications, vol. 52, no. 9, pp. 34-40, 2012.
[15] A. Veselinova and N. Ristova, "Using activities in moodle for development the language skill writing," *Journal of Process Management - New Technologies*, vol. 4, pp. 745-750, 2014.
[16] A. Veselinova, "Designing final tasks and acquiring the communicative competence with the task-oriented approach in the business english classroom," *The Journal of Teaching English for Specific and Academic Purposes*, vol. 4, no. 3, pp. 561-572, 2017.
[17] S. Ghosh, S. Biswas, D. Sarkar and P. P. Sarkar, "Association rule mining algorithms and genetic algorithm: A comparative study," *IEEE International Conference on Emerging Applications of Information Technology (EAIT)*, 2012, pp. 202-205.