Data Preprocessing of FRBCS for Early Warning of Student Learning

Qun Zhao [1a]、Jin-Long Wang [2b]、Pei-Chen Hung [2c*]、Shu-Yuan Chuang[2d]

[1] College of Science & Technology, Ningbo University, Ningbo, China
[2] Ming Chuan University, Taipei, Taiwan, R.O.C.
[a]zhaogun1981@163.com, [b]jlwang@mail.mcu.edu.tw, [c]laura@mail.mcu.edu.tw, [d]sharon@mail.mcu.edu.tw
Corresponding author: Pei-Chen Hung, Phone number: +886-2-28824564, E-mail: laura@mail.mcu.edu.tw

Abstract

This paper uses the log data in Moodle system to predict students' learning performance at the early stage of a semester. Since the data quality has great influence on the prediction accuracy, the Normal transformation and the Z transformation are utilized in the preprocessing phase. Then, the Fuzzy Rule-Based Classification System (FRBCS) is employed to create prediction model. The experiment results illustrate that data with Normal distribution can offer the higher prediction accuracy than other methods.

Key words: Learning warning, Prediction, Data preprocessing, Fuzzy rule-based classification system

Introduction

In the past, most universities paid great attentions to the mid-term warning for discovering the students with poor learning performance, who will be provided with the remedial courses and counseling methods for improving their learning status. However, after the mid-term examinations, even if these students are provided with additional learning supports, they cannot keep up with the pace of course progress. Therefore, this paper based on the log data of the Moodle system employs the technology of educational data mining (EDM) to predict the possible learning outcomes of students during the first 6 weeks of the semester. In this way, instructors can precisely obtain the students' learning status and offer learning supports at the early stage of a semester for achieving the target of early warning.

Data preprocessing is a very important task for educational data mining. Therefore, the student learning activity data should be processed by transformation methods in advance for enhancing the quality of data. The transformed data can then be applied to the appropriate data mining technique for model establishment and prediction. The whole data mining process involves a lot of preparatory work and planning process, in which the data pre-processing stage accounts for nearly 50% of the workload that is a critical work that cannot be neglected. In the process of data mining, as shown in Figure 1, the target is to transform the vast amount of data into useful information, to discover knowledge of learning behavior, to enhance student learning, and to complete the circular of continuous improvement [1][2].

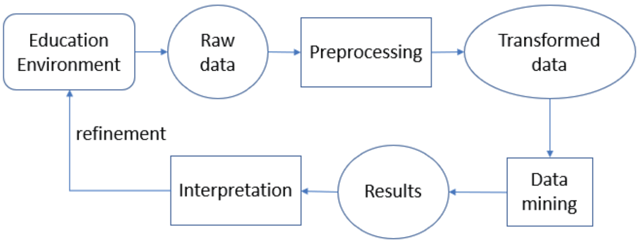

Figure 1. Educational data mining process
(Cristobal R. & Sebastian V., 2013)

In the data preprocessing stage, three common methods are used to refine the quality of data, including the processing of missing values, the processing of category data, and the scaling of data features. Since the range of each variable is very different to each other in the log data from the Moodle sysem, the standardization or transformation of the data is a very important step before the process of data mining. Therefore, this study uses the original, Normal distribution, and Z score methods to reconstruct the student learning activity data. After that, the re-formed data is classified and predicted by the fuzzy classification system to identify which method can have the better prediction accuracy. This study expects to predict the student's learning performance in the first 6 weeks of the semester, so that teachers can accurately acquire the student's learning outcomes at first-third of the semester for achieving the target of early warning, which can be used to offer early assistance for students with learning difficulty.

Research Methodology

Data mining is consisting of a lot of techniques for data collection, data preprocessing, data analysis and results reporting. In the phase of data collection, R language is used to access the log file of student activities from Moodle system directly through the RMySQL suite.

The data preprocessing mainly consists of three steps, including the retrieval of required fields in the log file, the normalization or transformation of each field, and the elimination of outlier data. As to the transformation, the data is cleaned and converted to the required format by three methods, including the original method, the Normal transformation, and the Z transformation. The original method indicates that the raw data is unchanged and remained to be the rudimentary information. The Normal transformation converts the original data into the new version matching the distribution of Normal. the Z transformation standardizes data based on the mean and the standard deviation.

https://doi.org/10.35745/ecei2019v2.108

Proposed method

In the past research, the classification methods were mainly used to establish a decision-making model based on the student learning activity log data and the final grade of the course, and to predict the learning outcomes of the students in the same course in the next semester. Among them, the final learning outcomes used in most studies are usually divided into pass and fail. However, by observing the log data of learning activities, the behaviors of the various learning activities of students falling on the borders between pass and fail are very similar.

If the training data of the model only uses the boundary between passing and failing to divide the learning behavior pattern, it will lead to relatively large misjudgment. Namely, the learning activities of pass close to the boundary are in fact very similar to the learning activities of fail close to the boundary. However, as the distance between the grades and the boundaries increases, the similarity between learning activities of pass and fail will gradually decrease. Therefore, this study attempts to use the fuzzy classification method, and hopes to improve the accuracy of the prediction in order to achieve the target of learning warning at medium-term.

The Mamdani model, proposed by Mamdani and Assilian, is constructed from the linguistic variables of the antecedents and outcomes of the rule, with multi-input and single-output (MISO) manner. The Mamdani model, shown in Figure 2, consists of four elements, including a fuzzifier, a knowledge base, an inference engine, and a defuzzifier. The fuzzification interface converts the input of the crisp values into the fuzzy values, and the knowledge base consists of a database and a rule base, in which the database includes fuzzy set definitions and parameters of the membership function. The rule base contains fuzzy IF-THEN rules and the inference engine, which utilizes suitable fuzzy rules to infer the fuzzy values. The defuzzification is converted the inferred results to the crisp value as the final results [3][4].

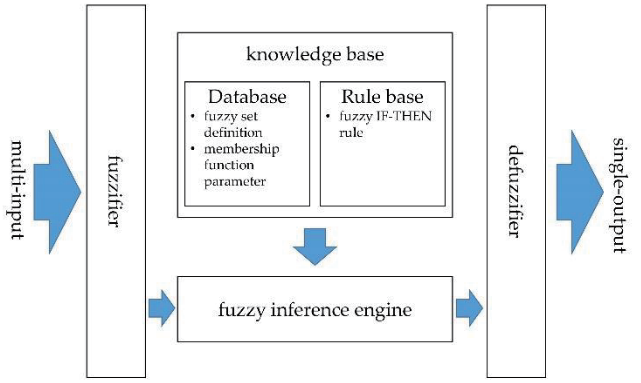

Figure 2. Mamdani model

Evaluation Results

In the Moodle system, there are a lot of learning activities that can be collected for further investigation, such as: view, post, reply, view of form, view of post, resource and upload. The target of prediction is to identify which students will pass or fail in the courses.

Ten courses are selected from the Moodle system in this study. The activity data from ten courses was processed by the three methods, including the original method, the normal transformation, and the Z score transformation. These three kinds of transformed data then are applied to the fuzzy classification method to predict the possible outcomes of students. With the comparison of three methods, it was found that the accuracy of the Normal distribution method can have the best performance, as shown in Figure 3 and Table 1. If the original data is used for the fuzzy classification system analysis, the prediction rate is no more than 10%. If the Normal distribution data is used, the prediction accuracy is higher than 90%, as shown in Figure 3. It is evidently that the transformation of learning activity data is a very important step for educational data mining.

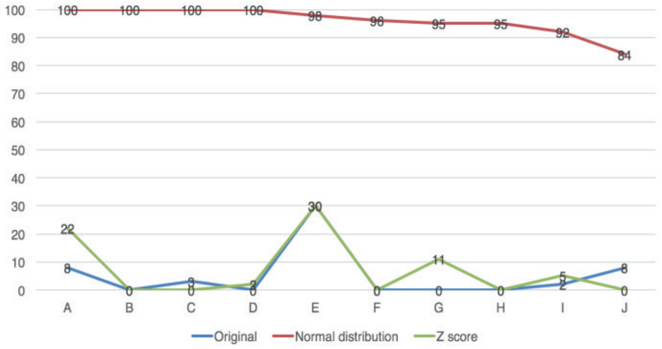

Figure 3. Comparison of three methods

Table 1. Comparison of three methods

course	Original	Normal transformation	Z transformation
A	8%	100%	22%
B	0%	100%	0%
C	3%	100%	0%
D	0%	100%	2%
E	30%	98%	30%
F	0%	96%	0%
G	0%	95%	11%
H	0%	95%	0%
I	2%	92%	5%
J	8%	84%	0%

Conclusions

In the process of data mining, pre-processing of data play an important role. The results illustrate that if the raw data is used to perform the prediction, the accuracy rate is about 5%, and the performance of the Normal transformation will be the best, which is up to 96%. Students whose grades fall on the boundary of pass and fail are quite similar in their behavior of learning, so it is very difficult to define the border with respective to the learning activities. Using the proposed fuzzy classification method, the boundary of pass and fail can be fuzzified. In this way, the prediction accuracy can be improved evidently.

References

[1] C. Romero and S. Ventura. *Data mining in education. WIREs Data Mining and Knowledge Discovery*, V.3, 2013, pp.12-27.

[2] C. Romero, J.R. Romero and S. Ventura. *Educational Data Mining: Applications and Trends,* Springer,2014, pp.29-64

[3] E.H. Mamdani. “Applications of Fuzzy Algorithm for Control a Simple Dynamic Plant.” *Proceedings of the Institution of Electrical Engineers*, Vol. 121, Issue 12, December 1974, pp.1585–1588.

[4] E.H. Mamdani and S. Assilian. “An Experiment in Linguistic Synthesis with a Fuzzy Logic Controller.” *International Journal of Man-Machine Studies*, Vol. 51, Issue 2, August 1999, p.135-147.