A GUIDE FOR ASSOCIATION RULE MINING IN MOODLE COURSE MANAGEMENT SYSTEM

Goran Avlijaš,
Milenko Heleta,
Radoslav Avlijaš

Singidunum University,
32 Danijelova Street, Belgrade, Serbia

Abstract:

The main goal of educational data mining is development and evaluation of different methods of exploration of educational data. Different analytical tools offer opportunities to analyze data generated by different kinds of Learning Management Systems (LMS). The goal of this paper is to describe the process of association rule mining in Moodle LMS, in theory and praxis, step by step. For practical demonstration we have used the data from one of the Singidunum University Moodle courses. The data was analyzed using the open source data mining software Weka. Selected variables included students' login frequency, number of accessed resources and forum messages, as well as the average performance on quizzes. With minor adaption of the proposed method any educator can realize benefits of association rule mining, regardless of prior knowledge in data analysis, and without having to purchase expensive software tools.

Key words:

data mining, moodle, association rule mining.

1. INTRODUCTION

The utilization of distance education systems has increased in the last decade, and virtual learning platforms are increasingly installed by educational institutions in order to offer e-learning options to students and enhance the quality of traditional courses. These e-learning systems provide different types of channels and work environments which enable distribution of information and correspondence between learners and educators. Among other features, these systems enable distribution of information to students, preparation of assignments and tests, engagement in discussions, collaborative learning within forums and chats *etc.*

Modular Object-Oriented Dynamic Learning Environment (Moodle) is one of the most commonly used open-source e-learning systems. It is an open-source LMS which enables development of comprehensive and flexible online courses and experiences (Rice, 2011). This system accumulates a large amount of data which can be used for the analysis of students' behavior. It registers every activity of a student, such as reading, writing, taking tests, performing various tasks and communicating with peers (Mostow and Beck, 2006).

Correspondence:
Goran Avlijaš

e-mail:
gavlijas@singidunum.ac.rs

DOI: 10.15308/Sinteza-2016-56-61

Most of the learning management systems include a database which contains different types of information: user data, academic performance, interaction data *etc.* Although some of the systems provide certain types of reporting, with large datasets it is difficult for a system itself to extract valuable information. Most of them do not provide built-in software which enables teachers to evaluate the course structure and content and its effectiveness. Therefore, a very promising ground for resolving this issue is the application of data mining tools (Zorrilla *et al.*, 2005).

2. EDUCATIONAL DATA MINING

Data mining (DM) represents automated recognition of implicit and interesting patterns from large amount of data (Klösgen, 2002). DM is an interdisciplinary scientific area which encompasses several computing paradigms: rule induction, decision tree, Bayesian learning, ANN, statistical algorithms, and so on. The most commonly used data mining techniques include classification, clustering, association rule mining, visualization, statistics, text mining *etc.*

In order to improve the learning process, researchers have investigated application of different data mining techniques in the educational context. Data mining tools allow teachers to analyze and visualize learning data in order to recognize useful patterns and evaluate the effectiveness of the course. The results obtained can be directed to students and certain activities or resources can be suggested that can improve the overall learning process.

According to Romero, Ventura, and García (2008), data mining in learning management systems represents an iterative process which should improve the overall learning and decision making process. Data mining process in educational context, as general data mining process, uses the following four steps:

- *Data collection*. While students use the system information is collected in the database. In case of Moodle, data is collected in a form of system logs.
- *Preprocessing*. After collection, the data needs to be transformed into the suitable format for analysis. A specific software can be used for preprocessing.
- *Data mining*. In order to develop a model and discover useful rules, the appropriate data mining algorithms should be applied at this stage.
- *Results evaluation*. Finally, educator interprets the obtained results and uses discovered knowledge to improve the learning and decision making process.

3. MINING ASSOCIATION RULES IN MOODLE

Association rule mining (ARM) has been used in learning management systems for finding correlations between items in order to: diagnose learning problems and offer students advice (Hwang *et al.*, 2003), determine suitability of learning materials (Markellou *et al.*, 2005), and identify patterns of performance disparity between groups of students (Minaei-Bidgoli *et al.*, 2004). ARM can also be used to: provide feedback to the teacher based on discovered relationships from students' usage information (Romero *et al.*, 2004), find most common errors (Merceron and Yacef, 2004) and optimise course content based on student interests (Ramli, 2005).

The Moodle learning management system is an open source learning platform, it represents an alternative to other similar commercial solutions. It is used in more than 240 countries worldwide by a large number of educational institutions, with more than 70 million users and more than 80,000 web sites (Moodle). Singidunum University implemented Moodle in 2006 and since then several thousand students have enrolled in more than 100 courses.

Moodle allows course administrators to access detailed logs of student activities, which keep track of materials and resources accessed by students and record every click for navigational purposes. Logs can be filtered by course, participant, day or type of activity. For some activities such as tests, a detailed analysis of each response is available in addition to the final score. Teachers can use logs to analyze students' performance, what and when they did something, and as such, they can be suitable for data mining.

Depending on the needs, users can choose between general/specific and commercial/open-source data mining tools. Open source machine learning software Weka (Waikato Environment for Knowledge Analysis) is developed by University of Waikato, New Zealand. WEKA comes with a set of data mining algorithms for data preprocessing, classification, regression, clustering, association rules and visualization (Witten *et al.*, 2011). In this paper, we used Weka because it is free, developed in Java and uses ARFF dataset external representation format.

Weka software enables recognition of interesting relationships between attributes in a form of association rules, which represent close correlation of support and

confidence. The confidence is defined as ratio of number of observations that contains the consequence and number of observations that contain the antecedent. The support is the ratio of the number of observation that contain both antecedent and consequence and the total number of observations in the dataset (Agrawal *et al.*, 1993).

In order to use a particular data mining algorithm, the dataset needs to be transformed into the appropriate format. Therefore, in order to apply the mining algorithm, preprocessing activities such as cleaning, transformation, enrichment, integration and reduction should be performed. Data preprocessing in Moodle is not that demanding as in other systems. However, activities such as data selection, acquisition, discretization and transformation have to be carried out. The following section describes data preprocessing and application of the selected association rule algorithm.

Data Preprocessing

The first step of data preprocessing is *data selection*. Although information is available for several thousand students in more than 100 different Moodle courses at Singidunum University, we have selected the course in Project Management. In the chosen course, students used different kinds of Moodle activities and resources (different types of resources, forum messages, quizzes). Table 1 shows the selected attributes. The total number of students in this study was 197.

Attribute	Description
stud_id	Student personal number
cs_view	Number of course visits
rs_view	Number of resources accessed
fr_view	Number of forum messages read
quiz_avg	Average score on quizzes

Table 1. Attributes used for each student

The next step of data preprocessing is *data acquisition*. The main prerequisite for data acquisition is creation of a Moodle log file which contains all necessary information. For the purpose of demonstration, we have generated a log table, which contains data about students' activities in the selected course. In order to create the final data set, the obtained data was transformed into a single summarized spreadsheet with the structure illustrated in Table 2.

stud_id	cs_view	rs_view	fr_view	quiz_avg
stud_1	82	33	5	23
stud_2	28	2	3	21
stud_3	40	15	3	28
stud_4	26	3	3	17
stud_n	60	47	8	19

Table 2. Structure of summarized dataset

The third step of data preprocessing is *data discretization*. In order to increase comprehensibility, numerical data needs to be discretized. In this process numerical values are divided into categorical classes that can be more easily understood by the teacher. We have discretized all the numerical values except student ID. For transforming continuous attributes into discrete attributes unsupervised global methods such as equal-width method, equal-frequency method or the manual method can be used (Dougherty *et al.*, 1995). As shown in Table 3, the equal-width method with three intervals and labels (low, medium and high) have been used for the selected attributes. The Weka system also enables discretization of numerical attributes using the 'discretize' filter.

stud_id	cs_view	rs_view	fr_view	quiz_avg
stud_1	high	low	low	high
stud_2	med	low	low	med
stud_3	high	high	low	low
stud_4	low	low	low	med
stud_n	med	med	low	high

Table 3. Sample of discretized data

The last step of data preprocessing is *data transformation*. After discretization, data needs to be transformed into the format required by data mining algorithm. In our case, Moodle log was exported as CSV file, which can be imported in WEKA and exported as Attribute-Relation File Format. An ARFF file is an ASCII text file that describes a list of instances sharing a set of attributes (Witten *et al.*, 2011). It is also possible to directly export Moodle log file as ARFF using specific tools for preprocessing such as Open DB Preprocess.

Application of association rule mining

Association rule mining is one of the most explored mining methods (Ceglar and Roddick, 2006). As defined by Agrawal, Imieliński, and Swami (1993), the problem of association rule mining is defined as: let $U = \{u_1, u_2, ..., u_m\}$ be a discrete universe, a finite set of objects. Let $A = \{a_1, a_2, ..., a_n\}$ be a finite set of attributes with binary values. Each object of universe U is described by attributes a_i, i = 1, 2, ..., n thus generating a data set. An associative rule is defined as an implication of the form $X \Rightarrow Y$ where $X, Y \in A$ and $X \cap Y \neq \emptyset$. The set of attributes X is called antecedent of the rule; the set of attributes Y is called consequent of the rule.

Although there are many $X \Rightarrow Y$ rules, in order to discover interesting ones, various measures of significance can be used; one of the most common are minimum thresholds on support and confidence. The support supp(X) is the proportion of objects in the data set which contain the attributes from X. The confidence of a rule is:

$$conf(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)} \quad (1)$$

These concepts are explained in the following example. For the data set given in Table 4 it is possible to infer some association rules, as well as the confidence and support parameters. For $X = \{A_1, A_2, A_3\}$ supp(X) =1/4 = 0.25 because there is one object (number four) for which there is a 'yes' value for every attribute. For example, the confidence of the rule $X \Rightarrow Y$, where X = {A1, A2} and Y = {A3} is:

$$conf(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)} = \frac{0.25}{0.5} = 0.5 \quad (2)$$

Attributes	A1	A2	A3
Instances	Student passed quiz	Student accessed platform	Student passed final exam
1	yes	yes	no
2	no	yes	yes
3	no	no	no
4	yes	yes	yes

Table 4. Simple example of data set

The previous rule $X \Rightarrow Y$ is interpreted as follows: the student who has passed the quizz and accessed the learning platform frequently has a 50/50 chance of passing the final exam. Apriori algorithm was the first algorithm that was used for association rules mining, and it served as the basis for other algorithms that were developed later, such as Apriori-TID, Eclat, FP-Growth, and so on (Ceglar and Roddick, 2006).

Several association rules algorithms can be used in the Weka, and for finding association rules we decided to use the Predictive Apriori algorithm. The Predictive Apriori is an improved version of the Apriori algorithm, which maximizes the probability of making an accurate prediction and resolves the issue of balance between support and confidence (Garcia *et al.*, 2011). Garcia, Romero, Ventura, and de Castro (2006) performed experimental tests on a Moodle course and confirmed better performance of Predictive Apriori in comparison to Apriori-type algorithm.

In the Predictive Apriori algorithm, measures of support and confidence are combined into a single value called predictive accuracy. This value is used to generate the Apriori association rule (Scheffer, 2001). In Weka software, Predictive Apriori generates „n" rules based on n value selected by the user. In the data set described, attribute selection is further complicated by the fact that some attribute values were not binary, which is why this involves a more extensive search.

The association rule generation is often the most appropriate technique due to several reasons:

- This is an exact method which excludes any subjective influence while analyzing a data set.
- The result is presented in a readable and easy-to-understand format.
- The number of generated rules can be limited in order to extract only the most accurate.
- It is expected that association rules have a great value when inferred from data set in the education domain because association rules can be treated as a hypothesis.

Results evaluation

The analysis was conducted in Weka in order to generate association rules. The Weka system requires that the attributes with numerical values do not participate in the association rule mining, which is why the were excluded from the analysis. We have executed the Predictive Apriori algorithm and generated rules with highest

predictive accuracy. Table 5 shows a list of the top 10 rules with highest predictive accuracy of the rule.

The number of discovered rule can be considerable with larger datasets. There can be many uninteresting rules, such as redundant rules, similar rules and rules with random relation. Rules relevant to educational purposes usually show the expected or conforming relationships or unexpected relationships. These rules can be very useful for teachers when making decisions about activities and identification of students with learning problems.

Table 5 shows that the number of read messages does not have a significant impact on the final grade outcome, since some of the top 10 rules are contradictory (such as rules 1 and 7). When it comes to login frequency (course view), according to rules 1, 4, 5 and 10, high quiz score cannot be expected when e-learning platform attendance is low. A similar case occurs with the number of accessed course resources (rules 1, 3 and 5). Rules 2, 7 and 8 show that if the course attendance is high, it is expected that the average quiz score obtained will be high. Using this information as a basis, the educator can give careful consideration to students who invest less energy in the learning process, since they are prone to failure. This way the educator can provide motivation and influence those students on time to pass the course.

	Best rules found	Predictive accuracy
1	cs_view=low fr_view=med rs_view=low 30 ==> quiz_avg=low 30	0.99077
2	quiz_avg=high fr_view=high 15 ==> cs_view=high 15	0.98141
3	cs_view=low rs_view=low 78 ==> quiz_avg=low 76	0.97457
4	cs_view=low fr_view=med 35 ==> quiz_avg=low 34	0.96928
5	cs_view=low fr_view=high rs_view=med 6 ==> quiz_avg=low 6	0.95017
6	cs_view=high fr_view=med rs_view=high 6 ==> quiz_avg=high 6	0.95017
7	cs_view=high fr_view=med 24 ==> quiz_avg=high 23	0.94704
8	cs_view=high rs_view=high 23 ==> quiz_avg=high 22	0.94410
9	quiz_avg=high rs_view=high 23 ==> cs_view=high 22	0.94410
10	cs_view=low 99 ==> quiz_avg=low 94	0.94095

Table 5. Predictive Apriori algorithm results

4. CONCLUSION

In this paper, we have described the association rule mining of Moodle data and how it can be utilized to enhance the learning process. As more students take part in online learning environments, databases expect to grow and mining opportunities expect to increase. In this sense, association rule mining can be useful for predicting student performance outcomes and identifying students who require special attention from teachers to increase the overall success ratio.

Data mining tools such as Weka are still too complex for most of the teachers and their capabilities go far beyond the average educator needs. In order to simplify application, educational data mining tools should be more user-friendly, intuitive, accompanied with proper visualization of results. This can be accomplished with integration of data mining tools and learning management system. Integrated tools could enable data processing in the same environment, so the responses and results can be directly implemented in the learning management system.

REFERENCES

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In ACM SIGMOD Record (Vol. 22, No. 2, pp. 207-216). DOI: 10.1145/170036.170072

Ceglar, A., and Roddick, J. F. (2006). Association mining. ACM Computing Surveys (CSUR), 38(2), 1-42. DOI:10.1145/1132956.1132958

Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In International Conference on Machine Learning, Tahoe City (pp. 194-202). DOI: 10.1016/b978-1-55860-377-6.50032-3

García, E., Romero, C., Ventura, S., & de Castro, C. (2006). Using rules discovery for the continuous improvement of e-learning courses. In Intelligent Data Engineering and Automated Learning–IDEAL 2006 (pp. 887-895). Springer Berlin Heidelberg. DOI: 10.1007/11875581_106

García, E., Romero, C., Ventura, S., de Castro, C., & Calders, T. (2011). Association rule mining in learning management systems. In C. Romero, S. Ventura, M. Pechenizkiy, & R.S.J. d. Baker (Eds.). Handbook of Educational Data Mining (pp. 93-106). CRC Press, Taylor and Francis Group, Boca Raton, FL. DOI: 10.1201/b10274-9

Hwang, G. J., Hsiao, C. L., & Tseng, J. C. (2003). A computer-assisted approach to diagnosing student learning problems in science courses. Journal of Information Science and Engineering, 19(2), 229-248.

Klösgen, W. (2002). Handbook of data mining and knowledge discovery. Oxford University Press, Inc. New York, NY, USA.

Markellou, P., Mousourouli, I., Spiros, S., & Tsakalidis, A. (2005). Using semantic web mining technologies for personalized e-learning experiences. Proceedings of the web-based education, 461-826, Grindelwald, Switzerland.

Merceron, A., & Yacef, K. (2004). Mining student data captured from a web-based tutoring tool: Initial exploration and results. Journal of Interactive Learning Research, 15(4), 319-346.

Minaei-Bidgoli, B., Tan, P. N., & Punch, W. F. (2004). Mining interesting contrast rules for a web-based educational system. In International Conference on Machine Learning and Applications, 2004. Proceedings. 2004 (pp. 320-327). Louisville, Kentucky, USA. DOI: 10.1109/icmla.2004.1383530

Moodle, a free open source course management system for on line learning. <http://moodle.org/> (2014).

Mostow, J., & Beck, J. (2006). Some useful tactics to modify, map and mine data from intelligent tutors. Natural Language Engineering, 12(02), 195-208. DOI: 10.1017/s1351324906004153

Ramli, A. A. (2005). Web usage mining using apriori algorithm: UUM learning care portal case. In International Conference on Knowledge Management, Malaysia (pp. 1-19).

Rice, W. (2011). Moodle 2.0 E-Learning Course Development. Packt Publishing Ltd. Birmingham, UK.

Romero, C., Ventura, S., & De Bra, P. (2004). Knowledge discovery with genetic programming for providing feedback to courseware authors. User Modeling and User-Adapted Interaction: The Journal of Personalization Research, 14(5), 425-464. DOI: 10.1007/s11257-004-7961-2

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. Computers & Education, 51(1), 368-384. DOI: 10.1016/j.compedu.2007.05.016

Scheffer, T. (2001). Finding association rules that trade support optimally against confidence. Principles of Data Mining and Knowledge Discovery (pp. 424-435). Springer Berlin Heidelberg. DOI: 10.1007/3-540-44794-6_35

Witten I.H., Frank E., & Hall M.A. (2011). Data mining: practical machine learning tools and techniques (3rd ed.). Morgan Kaufmann. New York, USA.

Zorrilla, M. E., Menasalvas, E., Marin, D., Mora, E., & Segovia, J. (2005). Web usage mining project for improving web-based learning sites. In Computer Aided Systems Theory–EUROCAST 2005 (pp. 205-210). Springer Berlin Heidelberg. DOI: 10.1007/11556985_26