Clustering Learners in a Learning Management System to Provide Adaptivity

Neslihan Ademi[1[0000-0002-9510-8182]] and Suzana Loshkovska[2[0000-0001-7630-7701]]

[1] International Balkan University, Skopje, North Macedonia
[2] Ss. Cyril and Methodius University, Skopje, North Macedonia
neslihan@ibu.edu.mk
suzana.loshkovska@finki.ukim.mk

Abstract. Learning Management Systems are a great source of data about the learners and their learning behavior. Educational Data Mining (EDM) together with Learning Analytics (LA) are emerging topics because of the huge amount of educational data coming from these systems. Knowledge gained from LA and EDM can be used for the adaptivity of learning systems to provide learners a personalized learning environment. This paper presents the clustering analysis of Moodle data in terms of learners' preferences on different assessment methods. Clustering is made by using four different algorithms and different number of clusters to find the most suitable method for a future adaptive learning system.

Keywords: Educational data mining · Learning analytics · LMS · E-learning · Log analysis · Clustering

1 Introduction

In formal and informal learning settings Learning Management Systems (LMSs) have been very popular as they support teaching and learning activities. They became a part of main stream education as supportive tools. With the capability of storing learners' data they give the opportunity to the researchers to extract knowledge from the educational data. The specific data mining research areas dealing with educational data are Learning Analytics (LA) and Educational Data Mining (EDM). In [23], authors quote the definition of Learning Analytics by The Society for Learning Analytics Research as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the learning environments". Educational Data Mining (EDM) uses machine learning methods to derive educational decisions [22]. Large quantities of available data and advances in machine learning techniques drive these areas to emerge.

Learning Management Systems (LMSs) offer opportunities for educational institutions to manage the learning process. Adaptive Learning Systems (ALSs) provide individualized content for the learners applying data analysis and machine learning techniques, but they are usually designed for specific purposes. LMSs are already capable to track interaction data of the learner. Integrating the logic and the framework of ALSs into LMSs would provide a better learning environment. With the help of different LA and EDM methods, adaptivity can be integrated into LMSs. Figure 1 represents our general idea and possible data flow for this integration.

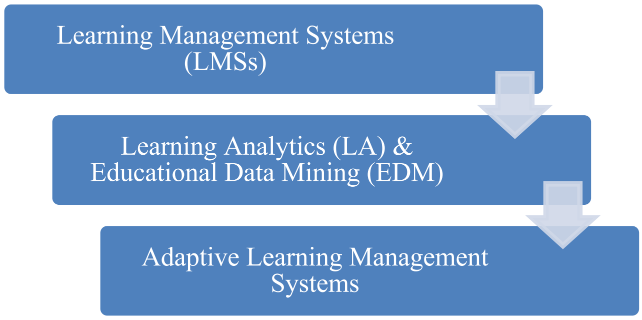

Fig. 1. Data flow for the adaptivity of LMSs

The integration of adaptivity into LMS requires firstly the collection of user data and this can be done through the system logs as we tried in our study. But it requires a big effort in terms of pre-processing to be able to extract details about users' online activities and preferences within the course. The second issue with providing adaptivity in a LMS is discovering trends in the data for decision making. Based on the preferences and online activities of the students, they can be grouped and different types of learning materials or assessment methods can be offered to the different user groups to increase their engagement and success. Clustering is one of the methods for grouping similar users.

In our study we partitioned log data by user characteristics and we tried different clustering methods to group the learners according to their behaviors in the learning system such as preferences of assessment activities and number of these activities. The purpose of the study is to find the best clustering method which can identify the differences of behavior and preferences between the learner groups. So these groups can be used in adaptation of the learning system to provide different activities to the different groups of learners. For example, students stacked to reproductive knowledge can be provoked to give effort on practical assessment methods.

The paper presents the analysis of Moodle log data by using clustering methods for the purpose of grouping the learners according to their assessment preferences. The second section defines the related work, third section explains the teaching process, the fourth section defines the used methodology for the analysis, while the fifth section gives results and discussion, and finally the last section is the conclusion.

2 Related Work

User modeling and understanding learners' behavior are important parts of adaptive learning systems. In user modeling there are two basic questions to be addressed; the first one is how to initialize the new user model and the second is how this model will be updated. Generally, this process involves diagnosis, classification, and control of the user parameters or characteristics.

Modeling learning-related attributes of students is essential in adaptive learning environments for the adaptation and feedback mechanisms [3]. Novel adaptive learning systems introduce the new development and deployment of adaptive mechanisms by using different attributes for user modeling. The structure of user models constantly changes over time. Some of them consider learning styles [5, 7, 10, 16, 20, 24] and some uses hybrid models [2] to present user characteristics. These user modeling approaches are mainly designed for adaptive learning systems which are field specific. So they are not suitable for a general framework which can be applied on any kind of subject area.

The data collected from users can be categorized like in [2], as demographical data through the registration process, explicit ratings for a subset of the available items, and implicit data from the user's online behavior. In this study we use the implicit data from the Moodle user logs to be able to discover online preferences of the students.

One of the most commonly used LMS is Moodle (modular object-oriented developmental learning environment), with a huge amount of data related to the students. When students are accessing Moodle, they use their personal account and digital profile is created for each student. All activities performed by the students are saved in log files. This data gives a big opportunity to analyze students' behavior and understand the characteristics of students.

Collected log data provides a descriptive overview of human behavior and insights about how people interact with existing systems. Generally, in observational log studies log data is partitioned by time or by user. Partitioning by time helps us to understand significant temporal features, such as periodicities and sharp changes in behavior during important events. It is also interesting to partition log data by user characteristics.[6]

Recently there are many studies in the literature about log analysis in e-learning environments [4, 8, 9, 12, 13, 18, 19, 21]. Some of the studies use learning analytics for monitoring students' online participation [18], some of them for profiling the students [1], and some of them for the grade prediction [9]. Most of them mainly focus on monitoring and visualization of the user activities within the learning system and their results are not used in decision making or in any adaptation process.

3 Teaching and Learning Settings

This section defines the organization of assessment methods within the Moodle course "User Interfaces" subject to this study. The reason of selecting this course is the wide range of activities given to the students for the assessment purpose. Table 1 gives those assessment methods and techniques used in the course based on the classification given in [11].

The teaching process was designed as blended learning which contains classroom teaching with the support of Moodle LMS. Students were given transferable points for any activity within the course, so that no effort was lost and students had freedom to select which activity to perform. This study aims to evaluate their preferences and possibility of grouping the students according to these.

Assessment is the main part of educational process, with teaching and learning. In our case, different assessment methods were used in a combination. Beside the tests for grading; different tools are used such as quizzes for assessing reproductive knowledge, practical assignments and projects for assessing practical knowledge.

Grouping students allows identifying excellent students or students who are lagging and need help. In other way students can be identified as good at practical work or good at reproductive knowledge.

Table 1. Assessment methods and techniques used in the course

Assessment Method and technique	**Nature / Purpose**	**Tools used**
Written assessment	Written responses based on individual experience Assesses prior knowledge	Partial exams and Final exam
Practical assessment	Results based on individual or team experience Assesses prior skills	Lab exercises
Homework assessment	Can be based on individual or team experience Assesses prior knowledge	Assignments Informal assignments
Assessment through classwork	Can be based on individual or team experience Assesses prior knowledge	In-class exercises
Project	Can be based on individual or team experience Assesses knowledge, skills and attitudes	Project or seminar work
Self-evaluation	Results based on individual experience Assesses prior knowledge	Self-evaluation tests
Reciprocal assessment – Peer assessment	Results based on individual experience Assesses prior knowledge	Students' evaluation on their colleagues
The system of granting and calculating transferable points "nothing is lost, everything is transferable"	Based on individual experience and on experience of teamwork with the teacher Assesses knowledge, skills and attitudes	All above

Some assessment methods are based on measuring existing knowledge; some are requiring creativity together with the knowledge. According to Montouri, reproductive learning is a form of education based on rote memorization and reproduction of existing knowledge where critical and creative thinking are largely ignored; and it does not prepare the learner for the complex, unforeseen and rapidly changing world [15]. On the other side critical and creative thinking are the required skills for 21th century; especially in engineering education and computer science. There are assessment methods which measure the reproductive knowledge such as classical quizzes and tests. On the other side there are other assessment methods to encourage the learner into more creative way of learning by the help of different assignments and projects.

4 Methodology

In this section, the used methodology is explained step by step. For the study, log data from Moodle LMS is used. In the complete process we used two software; RStudio[25] and WEKA [26] which are both open source under GNU. RStudio is a development interface for R language which allows easy manipulation of data. WEKA is useful in terms of its ready interface which contains machine learning algorithms for many data mining tasks.

For the online part of the blended classroom model, students were engaged in the activities of a Moodle course to enhance their classroom learning.

Moodle course activity logs include 58 different events or actions which belong to different types of users. All events and actions are grouped in four main categories such as view, add, update, and delete. Most of the actions are relevant to students' learning experiences. Moodle records every action done by students. It's possible to export log files in various file formats. In our study they are extracted in .csv format.

4.1 Data Collection and Pre-processing

For the study, log files are extracted from Moodle in .csv format and contained all activities of the students at the academic year 2018-2019. Log files belong to third year undergraduate students enrolled to User Interfaces course from different departments of Faculty of Computer Science and Engineering at the University of Ss. Cyril and Methodius in Skopje. The retrieved data was composed of approximately 250K rows which belong to 245 students.

Several pre-processing steps are applied to the log files, to keep on relevant and correct information. The first step was to convert all Cyrillic letters found in the original log files into Latin letters, as they are not recognized by R packages. The second step was to remove the actions logged by instructors and administrators selectively by filtering them using sqldf package in RStudio, as we want to analyze only the students' actions. Log data produced by the system is also removed by filtering the data where component field is system. Required fields are extracted and duplicate records are removed. We focused on the different types of assignments submitted through the system, quiz submissions, lab exercises, and self-evaluation quizzes by extracting Event and Event Context fields from the log files. All these activities are counted for each student, and these numbers of activities are normalized by using z-normalization to make the number of different activities comparable.

4.2 Clustering

We performed preliminary clustering experiments using the attributes extracted from the log data. The aim of the clustering is to group students exhibiting similar usage pattern in selection of assessment methods. All our clustering experiments were performed using WEKA to show the number of clusters and how many instances each cluster contains.

We have used four different clustering algorithms:

1. Expectation–Maximization (EM) algorithm generates probabilistic descriptions of the clusters in terms of mean and standard deviation for the numeric attributes and value counts for the nominal ones.
2. Simple K-means is an iterative algorithm that tries to partition the dataset into K number of pre-defined distinct non-overlapping clusters where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible.
3. X-means clustering [17] is optimized version of simple K-means algorithm where number of clusters are determined by the algorithm.
4. Density Based clustering [14] expresses the similarity between two fuzzy objects by distance probability functions.

4.3 Evaluation of Clustering Methods

In WEKA four different cluster evaluation modes are available:

1. Use training set (default): After generating the clustering WEKA classifies the training instances into clusters according to the cluster representation and computes the percentage of instances falling in each cluster.
2. Supplied test set: WEKA can evaluate clustering on separate test data if the cluster representation is probabilistic
3. Percentage split: WEKA can evaluate clustering on separate test data if the cluster representation is probabilistic (e.g. for EM).
4. Classes to clusters evaluation: In this mode WEKA first ignores the class attribute and generates the clustering. Then during the test phase it assigns classes to the clusters, based on the majority value of the class attribute within each cluster. Then it computes the classification error, based on this assignment and also shows the corresponding confusion matrix.

In our study we used the first option and then in addition, to support the clusters we used MANOVA (Multivariate ANOVA) test. MANOVA test was performed to determine if there were significant differences among the assigned clusters.

5 Results and Discussion

5.1 Clustering

Table 2 gives the mean and standard deviations as a result of EM clustering, when Fig. 2 gives the cluster means. Log likelihood is 4.43. As we can see from the Fig.2; students are grouped into 3 clusters. Cluster0 has 110 members and they are more active in self-evaluation quizzes which assess reproductive knowledge, so we can call them reproductive learners. Cluster1 has 15 members and they all have very low number of activities. Cluster2 has 121 members and all of them are very active in any kind of activities, they are eager to accomplish any kind of given task. By using this algorithm users can be categorized as

1. Highly active learners (Cluster2) – most probably they would not need any interruption by the instructor or by the adaptive system as they have effort on all types of activities.
2. Reproductive learners (Cluster0) – Mostly engaged to self-evaluation quizzes, they could be directed into more practical activities.
3. Inactive learners (Cluster1) – They are in need of motivation on any of the activities.

Table 2. Clusters EM (Log likelihood: 4.43094)

Attribute	Cluster 0 (n=110)		Cluster 1 (n=15)		Cluster2 (n=121)	
	M	SD	M	SD	M	SD
Formal Assignments	0.43	0.26	0.12	0.15	0.78	0.22
Lab Exercises	0.66	0.31	0.29	0.34	0.95	0.10
Optional Assignments	0.57	0.35	0.21	0.18	0.93	0.12
Self-Evaluation Quizzes	0.83	0.20	0.29	0.25	0.98	0.06
Quizzes	0.63	0.13	0.24	0.14	0.87	0.06

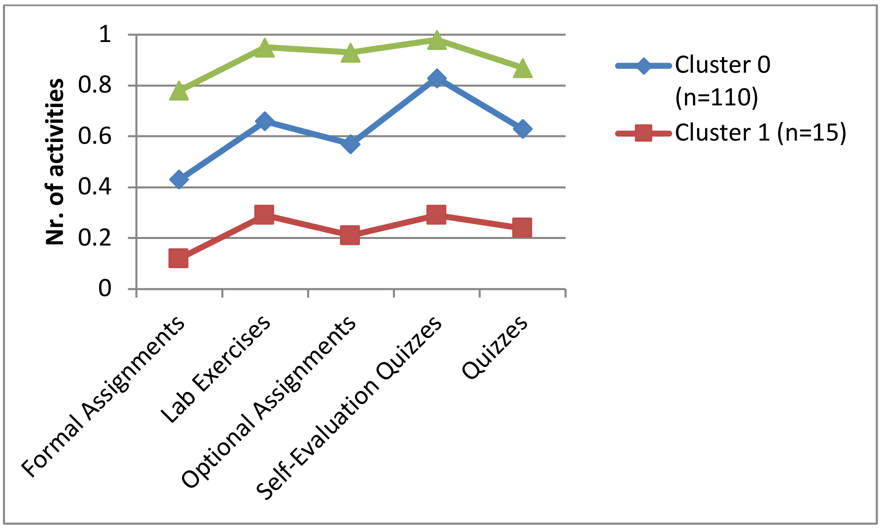

Fig. 2. EM Clustering Means

Figure 3 shows the clustering results with X-Means algorithm. X-means also determined 3 clusters from the data set. But in this case Cluster0 has 111 members which are very active in any kind of activities. Cluster1 has 80 members, mostly active in self-evaluation quizzes. This method did not provide significant difference between Clusters 0 and 1. Cluster2 has 55 members which are not so active.

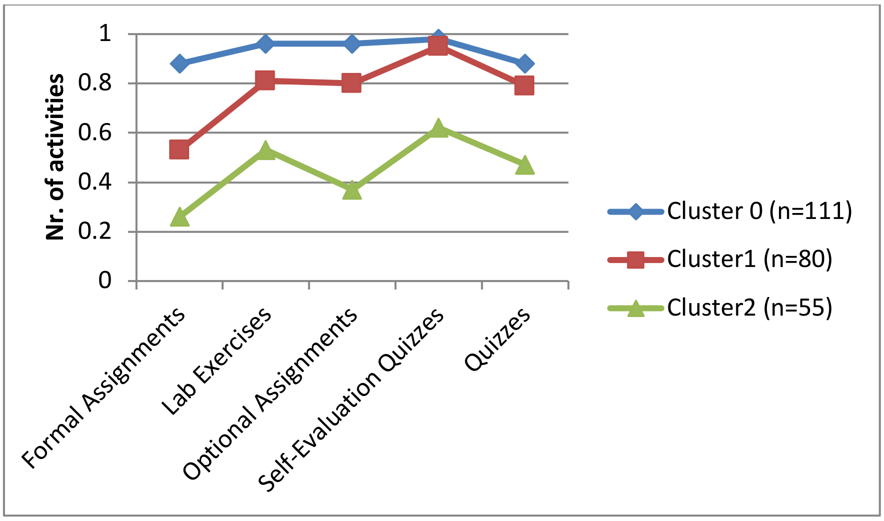

Fig. 3. X-Means Clustering Means

Figure 4 shows the results of simple K-means algorithm. We tried K-means algorithm firstly with 3 clusters. Cluster0 has 37 members and it is seen their lab exercises which represents the practical work is very low. Cluster1 has 44 members and they are active in lab exercises and self-evaluation quizzes. Cluster2 has 165 members and they are highly active in all kind of activities.

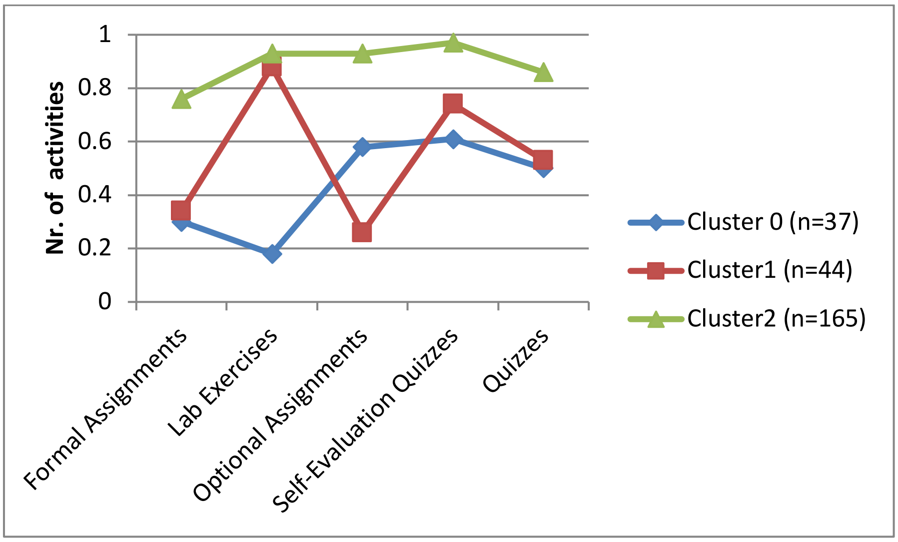

Fig. 4. Simple K-means Clustering means

Figure 5 shows the cluster assignments by density based clustering algorithm. With this algorithm we obtain two clusters. Cluster0 with 179 members shows the most engaged students which did most of the offered activities. Cluster1 with 67 members shows the students which have fewer activities.

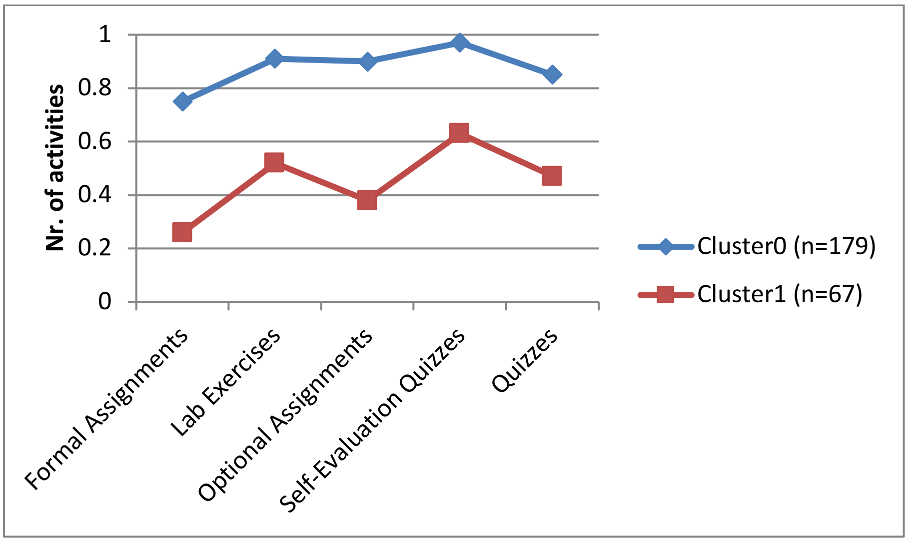

Fig. 5. Density based clustering centroids

5.2 Evaluation of the Clusters

Selected clustering algorithms EM, X-means and density based clustering define the number of clusters by themselves. EM and X-means clustering algorithms gave the same number of clusters which is three; while Density based clustering algorithm gave two. From the selected algorithms, only when using K-means algorithm number of clusters must be predefined.

In WEKA, we tried K-means algorithm with different k values and in each case WEKA generates the error parameter which is the distance between data points and their assigned clusters' centroids. We can pick k value at the spot where error starts to flatten out and forming an elbow. Figure 6 shows the application of elbow method in determining a good k number of clusters based on the sum of squared error. When k=6 error starts to flatten. So we can say that k-means would give the best result with 6 clusters.

For the evaluation of significance of the selected features we used MANOVA test. Table 3 shows p values as results of MANOVA test, where α=0.05. As it can be seen from the table in all cases the p-value is less than the alpha value, significance level is satisfied and null hypothesis is rejected.

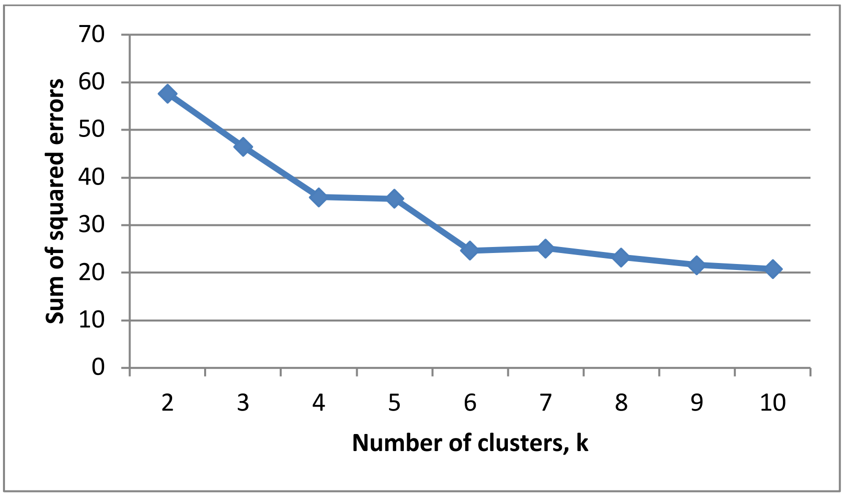

Fig. 6. K-means clustering error with different k values

Table 3. MANOVA test results (p values for each parameter)

	EM	K-means with 3 clusters	X-Means	Density based clustering
Formal Assignments	1.68E-35	1.36E-30	1.16E-64	2.3906E-31
Lab Exercises	1.3E-26	4.01E-78	4.28E-21	6.0679E-22
Optional Assignments	4.68E-32	3.17E-60	5.33E-51	6.4315E-44
Self-Evaluation Quizzes	2.97E-53	1.55E-23	2.94E-37	5.0215E-27
Quizzes	1.46E-69	5.46E-47	4.61E-62	3.4293E-49

6 Conclusion

Based on the focus of the tutor or the mechanism which will be used in decision making in adaptivity; different clustering algorithms could be selected to find the most suitable clustering method which fits the given data.

In this study we applied four different clustering algorithms to our data. We tried to identify which algorithm would be the paramount for our data and test the ideal number of clusters which determines user groups. EM, simple K-means, X-means and density based clustering were used as algorithms. As a result, EM and X-means clustering algorithms grouped the users in three clusters, while Density based clustering algorithm in two. We had to predefine number of clusters when using K-means algorithm, and this practice did not provide us with the finest clustering result, by itself.

We tried this algorithm with different number of clusters and used elbow method to see the flattening in error squares. Flattening occurred with six clusters and these clusters were not defining the user groups as we aimed. At the end, we noticed that K-means algorithm is unpractical for our purpose as it requires more analysis. EM was suitable at determining differences in user groups as it assigned the most critical inactive students as a separate group. X-means and Density based clustering did offer a big distinction between students which were very active or inactive/critical.

By using these algorithms, it appeared that learners can be grouped as reproductive learners or practical/creative learners, in general. This outcome will be used for grouping the learners in a future adaptive learning system based on user preferences of online activities as assessment methods.

In this study we limited our experiment to data of one course, as it offered more options in terms of different assessment methods. As a result we identified the most suitable clustering algorithms to our data. In the future studies, by applying the defined algorithms we will expand our research and integrate other details of student activities from the log files into the decision making process with an aim to find the best method for future adaptive learning system.

References

1. Akçapınar, G.: Profiling Students' Approaches to Learning through Moodle Logs. Proceedings of Multidisciplinary Academic Conference on Education, Teaching and E-learning in Prague 2015, Czech Republic (MAC-ETeL 2015). December, 7 (2015). https://doi.org/10.1016/j.bmcl.2014.11.048
2. Al-Shamri, M.Y.H., Bharadwaj, K.K.: Fuzzy-genetic approach to recommender systems based on a novel hybrid user model. Expert Systems with Applications. 35, 3, 1386–1399 (2008). https://doi.org/10.1016/j.eswa.2007.08.016
3. Bouchet, F. et al.: Clustering and Profiling Students According to their Interactions with an Intelligent Tutoring System Fostering Self-Regulated Learning. JEDM | Journal of Educational Data Mining. 5, 1, 104–146 (2013). https://doi.org/10.5281/ZENODO.3554613
4. Cocea, M., Weibelzahl, S.: Log file analysis for disengagement detection in e-Learning environments. User Modeling and User-Adapted Interaction. (2009). https://doi.org/10.1007/s11257-009-9065-5
5. Diaz, F.S. et al.: An Adaptive E-Learning Platform with VARK Learning Styles to Support the Learning of Object Orientation. In: 2018 IEEE World Engineering Education Conference (EDUNINE). pp. 1–6 IEEE (2018). https://doi.org/10.1109/EDUNINE.2018.8450990
6. Dumais, S. et al.: Understanding User Behavior Through Log Data and Analysis. In: Ways of Knowing in HCI. pp. 349–372 Springer New York, New York, NY (2014). https://doi.org/10.1007/978-1-4939-0378-8_14
7. Eltigani, Y. et al.: An approach to Adaptive E-Learning Hypermedia System based on Learning Styles (AEHS-LS): Implementation and evaluation. International Journal of Library and Information Science. 3, 1, 15–28 (2011)
8. Figueira, Á.: Mining Moodle Logs for Grade Prediction: A Methodology Walk-through. In:

Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality. pp. 44:1-44:8 (2017). https://doi.org/10.1145/3144826.3145394
9. Figueira, Á., Álvaro: Mining Moodle Logs for Grade Prediction. In: Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality - TEEM 2017. pp. 1–8 ACM Press, New York, New York, USA (2017). https://doi.org/10.1145/3144826.3145394
10. García, P. et al.: Evaluating Bayesian networks' precision for detecting students' learning styles. Computers and Education. 49, 3, 794–808 (2007). https://doi.org/10.1016/j.compedu.2005.11.017
11. Iancu, M.: The Assessment of Students' Competencies and Constructionism With Examples in Biological and Natural Sciences. In: Daniela, L. and Lytras, M. (eds.) Learning Strategies and Constructionism in Modern Education Settings. pp. 223–249 IGI Global (2018). https://doi.org/10.4018/978-1-5225-5430-1
12. Käser, T. et al.: Modeling exploration strategies to predict student performance within a learning environment and beyond. In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK '17. (2017). https://doi.org/10.1145/3027385.3027422
13. Konstantinidis, A., Grafton, C.: Using Excel Macros to Analyse Moodle Logs. UK Research.Moodle.Net. September, 4–6 (2013)
14. Kriegel, H. et al.: Density-based clustering. WIREs Data Mining and Knowledge Discovery. 1, 3, 231–240 (2011). https://doi.org/10.1002/widm.30
15. Montuori, A.: Reproductive Learning. In: Encyclopedia of the Sciences of Learning. pp. 2838–2840 Springer US (2012). https://doi.org/10.1007/978-1-4419-1428-6_811
16. Nafea, S.M. et al.: An Adaptive Learning Ontological Framework Based on Learning Styles and Teaching Strategies. September, 11–12 (2017)
17. Pelleg, D. et al.: X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In Proceedings Of The 17th International Conf. On Machine Learning. 727--734 (2000)
18. Poon, L.K.M. et al.: Learning Analytics for Monitoring Students Participation Online: Visualizing Navigational Patterns on Learning Management System. Presented at the (2017). https://doi.org/10.1007/978-3-319-59360-9_15
19. Rachel, V. et al.: Analytics on Moodle Data Using R Package for Enhanced Learning Management. (2018)
20. Rasheed, F., Wahid, A.: Learning Style Recognition: A Neural Network Approach. In: First International Conference on Artificial Intelligence and Cognitive Computing , Advances in Intelligent Systems and Computing, Springer. pp. 301–312 (2019). https://doi.org/10.1007/978-981-13-1580-0_29
21. Rebucas, R., Raga, R.C.: Analyzing students online learning behavior in blended courses using Moodle. Asian Association of Open Universities Journal. 12, 1, 1–20 (2017). https://doi.org/10.1108/AAOUJ-01-2017-0016
22. Romero, C., Ventura, S.: 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 (2010). https://doi.org/10.1109/TSMCC.2010.2053532
23. Siemens, G., Baker, R.S.J.D.: Learning analytics and educational data mining: Towards

communication and collaboration. In: ACM International Conference Proceeding Series. pp. 252–254 ACM Press, New York, New York, USA (2012). https://doi.org/10.1145/2330601.2330661
24. Truong, H.M.: Integrating learning styles and adaptive e-learning system: Current developments, problems and opportunities. Computers in Human Behavior. 55, 1185–1193 (2016). https://doi.org/10.1016/j.chb.2015.02.014
25. RStudio | Open source & professional software for data science teams - RStudio, https://rstudio.com/
26. Weka 3 - Data Mining with Open Source Machine Learning Software in Java, https://www.cs.waikato.ac.nz/ml/weka/