Classification, Visualization and Pattern Recognition using J48 and Zero R Machine Learning Algorithms

Ezekiel U Okike[1†] and Merapelo Mogorosi[2††]

University of Botswana, Gaborone, BOTSWANA

Abstract

. The identification and recognition of patterns is vital in Data mining and knowledge discovery. In the educational environment, mining data from Learning Management Systems (LMSs) data sets for useful strategic decision making has seen little investigation. Data was mined from Moodle LMS using the WEKA tool. J48 and Zero R ML algorithms selected from the WEKA tool were used to cluster, classify and visualize the data. The patterns of teaching and learning suggested that for clustering, both algorithms rated Quiz as the most used resource on the LMS resources followed by system logs which indicated that staff and students log on to the system to use the resources. In terms of classification, J48 is a better classifier than Zero R at 0.0, and 0.1 mean absolute errors, respectively. However, in terms of visualization of patterns, Zero R uses concentric colour coding while J48 uses tree format which becomes complex with large data sets. Therefore, Zero R is considered a better visualizer than J48. Overall, it was observed that there is significant correlation between students' use of LMS resources and academic performance in the sampled test case.

Key words:

Data mining, Machine learning, Learning management systems, J48 algorithm, Zero R algorithms.

1. Introduction

Data Mining (DM) has been defined as the process of discovering patterns in large data sets [1]. The essential steps in a data cycle process used to accumulate and store data involve defining a problem that needs a data mining approach, identifying pertinent data sources, collecting and storing data, data preprocessing, mining data, data post processing, knowledge discovery, learning and decision making, and a feedback process. The basic steps in a DM process were identified and explained in [2,3,4]. Looking for patterns in data helps to make sense of data in knowledge discovery [2,5,6,19]. The patterns discovered from data sets must be meaningful and useful for appropriate decisions to be made on its basis. One basic problem with large sets of data is that there will always be a growing gap between the generation of massive data sets and the complexity of understanding the data. Therefore, as the volume of data increases in the organization, the proportion of the data understood tends to decrease, although there exists lots of hidden knowledge from the data. This paper applied Data mining to a subset of massive data acquired by a university using teaching and learning management software such as Blackboard Learn or Moodle [1,4, 7, 8,9,10].

1.2 The Problem

Many African universities have accumulated data using Learning Management System (LMS) platforms such as Moodle and Blackboard. However, there is little or no investigations about applying data mining to LMS data sets for the purpose of discovering patterns in teaching and learning for strategic decision making in the African context. The present study sought to apply educational data mining to LMS data logs at an African university with the aim of discovering patterns of teaching and learning to provide information for useful strategic planning at the university.

1.3 Objectives of the study

The specific objectives of this study were:

i To investigate teaching and learning patterns from LMS data logs at an African University using pattern recognition tools from WEKA tool sets.

ii To evaluate the performance of J48 and ZeroR algorithms in classification, visualization, and pattern recognition

1.4 Research Questions

The following research questions were investigated in the study:

i. What were the recognizable patterns in teaching and learning from Moodle LMS logs for the sampled data set?

Manuscript received December 5, 2020
Manuscript revised December 25, 2020
https://doi.org/10.22937/IJCSNS.2020.20.12.25

ii. How does J48 and Zero R algorithms perform as classifiers, and visualizers?

2. Related Work

The usefulness and effectiveness of learning management systems in the academic environment have received much attention as shown in [2,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19]. The use of the Waikato Environment for Knowledge Analysis (WEKA) tool in Educational Data Mining (EDM) has also been demonstrated in a few scenarios such as described in [8,9]. A recent study in [2] applied the EDM concept in the African context with a view to discovering teaching and learning patterns from Moodle LMS. In addition, the study concluded that there is significant relationship between the use of LMS resources and student's academic performance.

The process to acquire and store data in organizations usually follow a cycle approach as explained in [2,20]. A Data Cycle (DC) in knowledge discovery may be illustrated in figure 1.

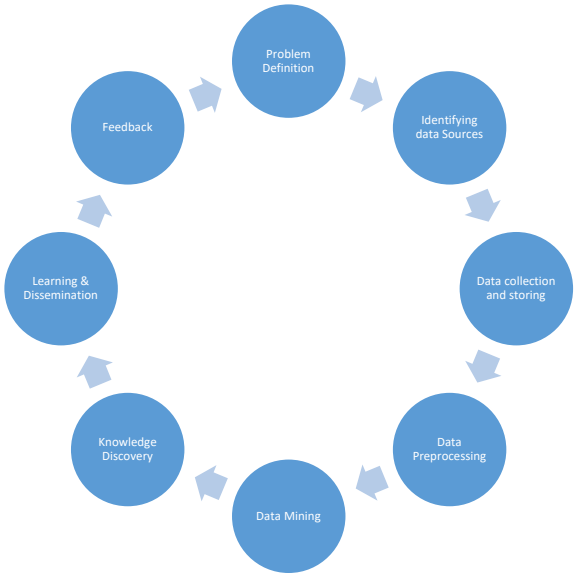

Figure 1 Data Cycle in Knowledge Discovery (DCKD) using Data Mining.

The process starts by defining a problem domain, identifying the sources of data in the domain, collecting, and storing data, mining, and processing the data to discover new knowledge, and finally disseminating the gained knowledge through appropriate mechanisms and also getting necessary feedback from users of the system.

3. The Empirical Study

Logs of Moodle LMS data were download as csv files and mined using a WEKA tool. The components of the logs included student's data, staff data, courses data and LMS resources data. In relational terms, the components of each of the data sets may be described as follows:

i. Students Data (surname, firstname, idno, logintime, system IP address)

ii. Staff Data (surname, firstname, staffed, logintime, system IP address)

iii. LMS Resources (System, Quiz, forum, chat, file submission, folder, Assignment)

J48 and Zero R algorithms from the WEKA tool were applied on the data through the stages of pre-processing, classification, clustering, and visualization.

4. Result and Discussion

4.1 Classification using J48 and Zero R

Fig. 2 and Fig.3 show the result of classifying the same course using both J48 and ZeroR algorithms. J48 gives higher number of correctly classified instances, and lower number of incorrectly classified instances. In addition, the error is 0.0. Therefore, J48 achieves better classification than ZeroR (research question 3).

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       29765               99.0813 %
Incorrectly Classified Instances       276                0.9187 %
Kappa statistic                          0.9872
Mean absolute error                      0.0026
Root mean squared error                  0.0387
Relative absolute error                  1.8404 %
Root relative squared error             14.4643 %
Total Number of Instances            30041
```

Figure 2 Result of classification using J48 course A

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       29765               99.0813 %
Incorrectly Classified Instances       276                0.9187 %
Kappa statistic                          0.9872
Mean absolute error                      0.0026
Root mean squared error                  0.0387
Relative absolute error                  1.8404 %
Root relative squared error             14.4643 %
Total Number of Instances            30041
```

Figure 3 Result of classification using ZeroR course A

4.2 Visualization using J48 and Zero R

Figs.4-6 show the visualized pattern of teaching and learning using Zero R in the selected courses. A plot matrix of four rows and four columns may be clearly recognized. The rows represent (event name, components, event context and user full name) and the columns have the same labels. The visualization presents cells representing activities. The results of Fig. 4 shows Quiz (light blue), System (red), Forum (blue), File (green), Assignment (pink), Folder (yellow), File submission (purple), and Chat (marron). The figure also shows that Quiz (light blue) appears more with light blue colours as the highest concentration, followed by System, Forum, File, Assignment, Folder, File Submission and Chat. This explains the pattern of how the lecturer taught the class (research question 1). The access of the course (System: red) led learners to use the quiz tool (light blue) more often because they had to discuss (Forum: blue) the concept of the given topic and read notes (File: green).

The learners also had to work on the given assignments (pink) presumably to check whether they had understood the concepts of the topic. Furthermore, an interesting feature in the plots is a straight line from the bottom left corner to the top right corner. The significance of this is that there is no interaction along the diagonal because the same parameters were being compared. The class colour for the various components is also visible as shown in bottom logs (royal blue), System (Red), Quiz (light blue), Forum (blue), Assignment (pink), File (green), Folder (yellow), File submission (purple) and Chat (maroon) (research question1).

4.3 Measuring the Performance of J48 and Zero R in Mean absolute error(MAE) and Mean Root Square Error (RMSE)

Considering fig. 2 and fig. 3, the lower value of measure of performance of J48 in terms of mean absolute error, and root mean squared error, as well as the higher number of correctly classified instances indicate better fit. The result was consistent in all course observed (research question 2).

Considering fig. 4, and fig. 5 Zero R appear to be good at visualization, as it uses colour coding to represent patterns visibly, unlike J48 which uses trees to represent patterns. However, the trees become cumbersome and complex with very large data sets. Tables 1,2, and 3 present a comparison of J48 and Zero R in detail (research question 2)

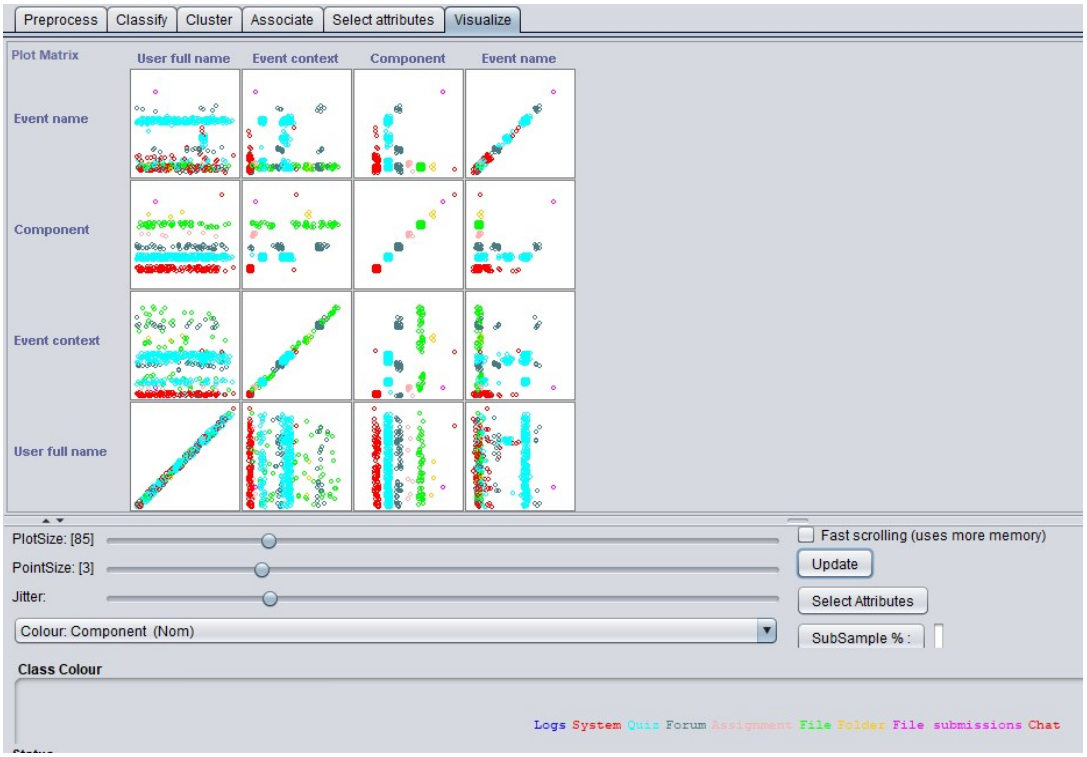

Figure 4 Course1 Plot Matrix

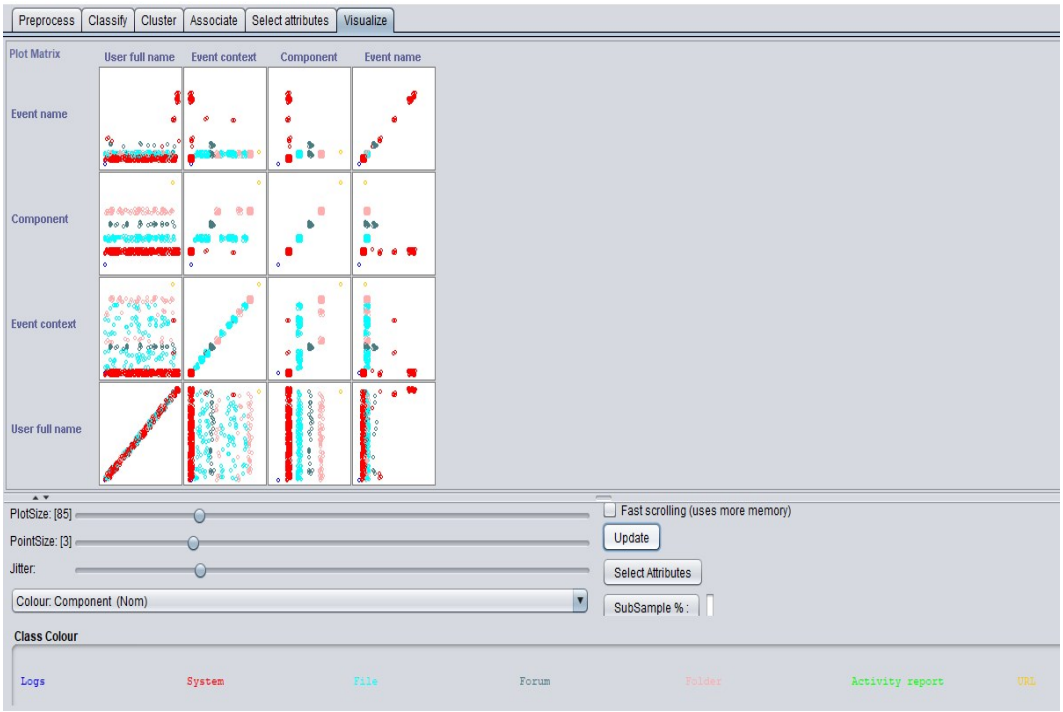

Figure 5 Course 2 Plot Matrix

Table1 Results of Zero R and J48 algorithms Classification

Activity	Course	J48	Zero R
Classification	Course 1	Mean absolute error:0.0028 Root Mean Square error: 0.039 Relative absolute error: 1.9191% Root absolute error:14.3847% Pattern: Quiz (38358), System (17910), Forum (8663), File (8566), Assignment (1235), Folder (514), File submission (172) and Chat (37)	Mean absolute error: 0.147 Root Mean Square error: 0.2711 Relative absolute error: 100% Root absolute error: 100% Pattern: Quiz (38530), System (18108), Forum (8802), File (8566), Assignment (1395), Folder (514), File submission (172) and Chat (122),
	Course2	Mean absolute error:0.0026 Root Mean Square error: 0.0387 Relative absolute error: 1.8404% Root absolute error:14.4643% Pattern: system (11920), quiz(8301), Forum (4476), File (4394), Assignment (257), Chat(247), Url (125) and File submission (39)	Mean absolute error: 0.1434 Root Mean Square error: 0.2678 Relative absolute error: 100% Root absolute error: 100% Pattern: System(12109), Quiz (8308), Forum (4485), File (4394), Assignment (309), Chat (270), URL(125), File submission (38) .
	Course3	Mean absolute error:0. Root Mean Square error: 0. Relative absolute error: 0% Root absolute error:0% Pattern: system (26220), File(1022), Folder (570), Forum (258), URL (2).	Mean absolute error: 0.1676 Root Mean Square error: 0.2894 Relative absolute error: 100% Root absolute error: 100% Pattern: System(2622), File (1022), Folder(570), URL (2) .

Table 2 Result of J48 and Zero R Clustering

Activity	Course	J48	ZeroR
Clustering	Course 1	Quiz	Quiz
	Course2	System	System
	Course 3	System	System

Table 3 Result of J48 and Zero R Visualization

Activity	Course	J48	ZeroR
Visualization	Course 1	Tree format	Colours. Dominant color for the highest class (light blue:Quiz)
	Course 2	Tree format	Colours. Dominant color for the highest class (red:system)
	Course 3	Tree	Colours. Dominant color for the highest class (red:system)

From Fig. 5 (course 2), the visualization presents cells representing activities. The results above show System (red), Quiz (torques blue), Forum (blue), File (orange), Assignment (green), Chat (pink), URL (purple) and File submission (maroon). The figure shows System (red) tool to be the highest concentration, followed by Quiz, Forum, File, Assignment, Chat, URL and File Submission. This explains the pattern of how the lecturer taught the class. The learners accessed the course (System: red) more often which led them to use the quiz tool (torques blue) to evaluate themselves on the course concepts, and they used discussion tool (Forum blue) to discuss the given topic, access directed links (URL: purple) and read course notes (File: orange). The learners also had to work on the given assignments (green) for the lecturer to check whether they had understood the concepts of the topic and submit (maroon) back their assignments for marking (research question1).

4.4 Comparison of Results

Table1 – Table 3 show the comparison of Zero R and J48 algorithms using Classification, Clustering and Visualization. For classification, J48 achieves better performance in terms of classifying higher number of instances at 0.0 error rate. Zero R achieves better visualization with colour coding than J48 which uses tree formats. Both algorithms predict Quiz and system as highest used LMS resources in all the courses investigated. A selected course was taught by a lecturer who showed enthusiasm in online activities on Moodle.

5. Conclusion

With reference to the research questions of this paper, findings indicated in the patterns of teaching and learning the substantial use of quiz (about 70%) for assessment purposes, followed by resource tools (file, folder and URL for posting notes and communications between lectures and students (research question 1). J48 achieves better performance as a classifier, while Zero R achieves better performance as a visualizer (research question 2)

References

[1] I. H. Witten and E. Frank, Data mining: practical machine learning tools and technology (2nd ed). New York: Elsevier, 2005.

[2] E. U. Okike and M. Mogorosi, "Educational data mining for monitoring and improving academic performance at university levels," *International Journal of Advanced Computer Science and Applications,* 11(11), 570-581, 2020.

[3] G. Mariscal, O. Marban and C. Fernandez, "A survey of data mining and knowledge discovery process models and methodologies," The Knowledge Engineering Review, 25(2), 137-166, 2010.

[4] O. Marban, G. Mariscal and J. Segovia, " A data mining and knowledge discovery process model," Open Access Data Base, 1-16. 2009.

[5] S.N. Hamade, "Students perceptions of learning management systems in a university environment: Yahoo Groups Vs Blackboard. Proceeding of the Ninth International Conference on Information Technology, New generations," 594-599, 2012.

[6] Y. B Kurata, R. M. I. P. Bano and M.C. Marcelo, " Effectiveness of learning management system application in the learnability of tertiary students in an undergraduate engineering program in the Philippines. In: Andre T. (eds) Advances in Human Factors in Training, Education, and Learning Sciences", AHFE 2017. *Advances in Intelligent Systems and Computing*, vol 596, Springer, Cham, 2017.

[7] Y. Ghiley, "Effectiveness of learning management systems in higher education: Views of lecturers with different levels of activity in LMS," Journal of Online Higher Education, 3(2), 29-50, 2019.

[8] C. Romero, S. Ventura, and E. Garcia, " Data mining in course management systems: Moodle case study and tutorial. Computers & Education. 51, 368-384, 2008.

[9] E. Garcia, C. Romero, S. Ventura and C. de Castro, " A collaborative educational association rule mining tool. The Internet and Higher Education, 14(2), 77-88, 2011.

[10] A.K. Alhazmi, and A. Rahman, "Why LMS failed to support student learning in higher education institutions," *IEEE Symposium on E-Learning, E-Management and E-Services, Kuala* Lumpur, 1-5, 2012.

[11] R.Babo, and R. Azevedo, Higher education institutions and learning management systems: Adoption and standardization. IGI Global, 2012.

[12] J. G. Boticario and O. C. Santos, Issues in developing adaptive learning management systems for higher education institutions. Retrieved November 5, 2020 from https://core.ac.uk/reader/55533720. .

[13] N. Darko-Adjei, Students perceptions and use of the Sakai learning management system in the university of Ghana. Retrieve November 5, 2020 from http://ugspace.ug.edu.gh/handle/123456789/26847.

[14] H. Coates, R. James and G.A. Baldwin, Critical examination of the effects of learning management systems on university teaching and learning," Tertiary Education Management **11,** 19–36, 2000.

[15] L. V. Ngeze, Learning management systems in higher learning Institutions in Tanzania: Analysis of students' attitudes and challenges towards the use of UDOM LMS in teaching and learning at the University of Dodoma, 2016. Retrieved October 2, 2020 from https://pdfs.semanticscholar.org/2c36/89101b4a64c85c25f3179c5a95e50f8a719a.pdf

[16] M. F. Paulsen, "Experiences with learning management systems in 113 European institutions," Journal of Educational Technology & Society, 6(4),.134-148, 2003.

[17] N. Fathema, D. M. Shannon and M. Ross, "Expanding the technology acceptance model (TAM) to examine faculty use of learning management systems (LMSs) in higher education institutions," Journal of Online Learning and Teaching, 11(2), 210-, 2015.

[18] V. Mhetre, "Classification based data mining algorithms to predict slow average and fast learners in educational systems using Weka," *ICCM*, 475-479, 2017.

[19] J. Talukdar, and S.K. Kalita, "Detection of breast cancer using data mining tool (Weka)," International Journal of scientific Engineering, 6(11), 1124-1128, 2015.

[20] N. Ahituv, "What should be taught in an academic

Ezekiel U. Okike received a B.Sc. and Ph.D. degrees in Computer Science, a Master of Information Science all from University of Ibadan, Nigeria. He is currently the cluster chair of Information Systems Cluster, Department of Computer Science, University of Botswana. He is a Senior member of IEEE, and a member of ACM. His research interests are in Information Systems, Software Engineering, Software Measurement, Software Quality, Machine Learning, Information Security/Cyber Security, E-government.

Mogorosi Merapelo is a graduate student in the Department of Computer Science, University of Botswana **.**