



Improving the portability of predicting students' performance models by using ontologies

Javier López-Zambrano^{1,3} · Juan A. Lara² · Cristóbal Romero³ 

Accepted: 9 March 2021 / Published online: 24 March 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

One of the main current challenges in Educational Data Mining and Learning Analytics is the portability or transferability of predictive models obtained for a particular course so that they can be applied to other different courses. To handle this challenge, one of the foremost problems is the models' excessive dependence on the low-level attributes used to train them, which reduces the models' portability. To solve this issue, the use of high-level attributes with more semantic meaning, such as ontologies, may be very useful. Along this line, we propose the utilization of an ontology that uses a taxonomy of actions that summarises students' interactions with the Moodle learning management system. We compare the results of this proposed approach against our previous results when we used low-level raw attributes obtained directly from Moodle logs. The results indicate that the use of the proposed ontology improves the portability of the models in terms of predictive accuracy. The main contribution of this paper is to show that the ontological models obtained in one source course can be applied to other different target courses with similar usage levels without losing prediction accuracy.

Keywords Educational data mining · Predictive modelling · Student performance · Transfer learning · Model portability · Ontology

✉ Cristóbal Romero
cromero@uco.es

Javier López-Zambrano
jlopez@outlook.com; jlopez@espam.edu.ec

Juan A. Lara
juanalfonso.lara@udima.es

¹ Escuela Superior Politécnica Agropecuaria de Manabí (ESPAM MFL), Faculty of Computing, SISCOM Group, 131106 Calceta, Ecuador

² Department of Computer Science, Madrid Open University, UDIMA, Carretera de La Coruña, KM. 38500, Vía de Servicio, No. 15, 28400 Collado Villalba, Madrid, Spain

³ Department of Computer Science and Numerical Analysis, University of Córdoba (UCO), 14071 Córdoba, Spain

Introduction

In recent decades, one of the main educational milestones is the advent of a new form of learning called e-learning (electronic learning), based on the use of the internet and technology to support students' online education. Nowadays, this form of learning is becoming particularly important due to the limitations defined by the authorities to restrain the spread of pandemics such as the one caused by Covid-19. The use of e-learning poses important advantages including the enabling of a more flexible temporal and spatial interaction than other forms of learning. Besides, vast amounts of learning process data can be collected, since it is based on the use of Learning Management Systems (LMS). Moodle (Dougiamas & Taylor, 2008) is one of the most used LMS overall, because, among other advantages, it is free, open and there is an important community of users who support its development. Data recorded by Moodle, in particular those that reflect students' interactions with educational resources, can be of great interest and applicability for building student behavior models. To analyze these data, approaches such as Educational Data Mining (EDM) and Learning Analytics (LA) are useful (Romero et al., 2008). In EDM, a field whose purpose is the extraction of knowledge from educational data, there are well-defined problems that have been addressed by the scientific community, such as the prediction of students' performance (Romero & Ventura, 2013, 2020). Recently, it is more frequent to find works that propose new approaches to analyzing educational data for a particular course. However, one of the due challenges is creating models for a particular course that can be useful when used in other courses (Baker, 2019). These are what we call transferable or portable models (Boyer & Veeramachaneni, 2015).

In our previous work (López-Zambrano et al., 2020), we obtained models generated from Moodle's logs data and we studied the degree of portability of the models between subjects, grouped by area of knowledge and by the usage level of platform resources. We used Moodle's native raw attributes which, in certain combinations of courses, led us to a certain loss in the portability of models since these low-level attributes are very dependent on each particular course. To overcome this limitation from our previous research, in this paper we present a new approach based on the use of resources from the semantic web area, in particular, ontologies (Fong et al., 2011; Tang & Fong, 2010). One of the most promising lines in this respect, particularly when analyzing logs of students' interactions with the LMS, is the categorization or taxonomy of attributes. In this regard, Bloom's taxonomy plays an important role. Bloom's taxonomy is a multi-tiered model of classifying thinking according to six cognitive levels of complexity which in this new version are: Remembering, Understanding, Applying, Analyzing, Evaluating and Creating (Forehand, 2005). Based on this idea, some works have even defined correspondence between the levels of Bloom's taxonomy and the different actions conducted by students in Moodle (Rollins, 2010). Some authors (Cerezo et al., 2020) proposed a categorization of low-level attributes into different higher-level codifications, such as Executing, Planning, Learning,

and Reviewing. Precisely, our research aims to evaluate the degree of portability of models built by using ontologies of interaction-with-the-platform attributes. To do so, we defined an ontology inspired by Bloom's taxonomy and based on the work by Cerezo et al. (2020), with the purpose of conducting a comprehensive study to measure the degree of portability of the models built based on that ontology (denoted as ontological models), compared with a previous similar study conducted by the authors López-Zambrano et al. (2020) in which we did not use ontologies but instead employed low-level Moodle attributes (denoted as non-ontological models). The models have been built from students' interactions with Moodle logs and the class attribute to predict is binary and represents whether or not the student will pass the course (Pass/Fail). In this work, the courses have been grouped according to the usage level of Moodle activities/resources. This approach has already been used in previous studies with satisfactory results (López-Zambrano et al., 2020). Taking all this into consideration, the global objective of this paper is to provide an answer for the research question below:

- Can the ontological models obtained in one (source) course be applied in other different (target) courses with a similar usage level without losing prediction accuracy?

The rest of the paper is organized as follows: Sect. “[Background](#)” reviews the literature related to this research. Section “[Materials and methods](#)” describes the data and the experiments. Section “[Results](#)” includes and discusses the results obtained. Finally, Sect. “[Conclusions](#)” presents the conclusions and future lines of research.

Background

Achieving generalizable and portable models is still an important challenge in the area of EDM, in spite of the important advances made in the last few years (Boyer & Veeramachaneni, 2015; Ding et al., 2019; Gašević et al., 2016; Hunt et al., 2017). In fact, Baker (2019) has considered what he calls the “Generalizability” or “New York City and Marfa” problem as one of the main challenges for the future of EDM, which is explained in detail in López-Zambrano et al. (2020).

To address this challenge, the resource of resources from the semantic web seems to be a promising line. The semantic web is an extension of the current web in which information is provided with a certain meaning, which makes cooperation and portability easier (Dhuria & Chawla, 2014). Fundamental resources from the semantic web are the ontologies, because they provide a common understanding of a domain. In particular, they may be interesting resources in the e-learning field (Al-Yahya et al., 2015).

In this regard, several particular works should be highlighted. In Octaviani et al. (2015) they present a tool, called RDB2Onto, for creating ontologies from Moodle logs, but this work does not validate the utility of such an ontology. In Castro and Alonso (2011) they propose a general architecture for EDM in which there is an educational ontology, but they do not define or develop the ontology, only providing

a general statement of it as a part of a higher-level architecture. There are even some works such as the one presented in Chang et al. (2020) where they utilize data mining techniques (association in this case) to build ontology-driven tutoring models for intelligent tutoring systems (this is precisely the opposite process to ours since we use the ontology for a further data mining analysis).

These previous works present general approaches. Other more specific works bear greater similarity to our study because they define particular ontologies to facilitate the EDM process. We found some works where the ontology created is not focused on attributes of students' interaction with the LMS. In Marinho et al. (2010) they propose an ontology to model EDM tasks, techniques, and parameters. In Grivokostopoulou et al. (2014) they propose an educational system that utilizes ontologies and semantic rules to enhance the quality of educational content (curriculum) and the learning activities delivered to each student. In Noura et al. (2019), they propose an ontological model for assessment analytics. And finally; in Dorça et al. (2017), they present an approach for the automatic and dynamic analysis of learning object repositories in which ontology models the relationships between the attributes and learning styles of the learning objects.

Other related works are those that define ontologies to model data of students' interactions with LMS resources. In El-Rady (2020), they propose an ontology where the student is the main class from which a series of associations arise that are connected to other classes that model the students' data (education, profile, social activities, etc.). That ontology is used as a part of a validation process to predict student dropout rates. Other related works are based on the idea of organizing the interaction attributes as part of a kind of taxonomy. It is worth highlighting the work presented in Cerezo et al. (2020), where they propose a process mining method for a self-regulated learning assessment, and make use of an ontology inspired by Bloom's taxonomy. In Montenegro-Marin et al. (2011), they also propose an ontology based on the idea of taxonomy, but not restricted to interaction attributes, as they consider many other features, such as the curriculum design, productivity, management, and so on. However, they do not validate the utility of the ontology.

Considering all the previous works, and to the best of the authors' knowledge, our work presented in this paper is the first that analyses the power of ontologies as a resource that makes the portability of EDM models easier and, in particular, it is also the only one for that purpose which is based on data from the students' interactions with the LMS. Furthermore, it is the first research that depicts a comparative study against a previous non-ontological similar approach, with the purpose of demonstrating the performance improvement obtained when using ontologies. Both of these innovative aspects are the core contributions of this paper.

Materials and methods

In this section, we describe both the data used and the preprocessing tasks we applied to them in order to transform the raw data gathered from the Moodle logs to the high-level attributes of the proposed ontology. We also describe the experimentation that we carried out in order to address our research question.

Table 1 Information of all subjects

Subject	Code	Degree	Year	#Users	Moodle usage
Introduction to programming (group 1)	IP1	Computer	1	144	Medium
Introduction to programming (group 2)	IP2	Computer	1	145	High
Programming methodology (group 1)	PM1	Computer	1	114	Medium
Programming methodology (group 2)	PM2	Computer	1	119	High
Professional computer tools	PCT	Computer	1	124	Medium
Databases	DB	Computer	2	58	Medium
Human computer interfaces	HCI	Computer	2	260	High
Information systems	InS	Computer	2	188	Medium
Software engineering	SE	Computer	2	58	Medium
Interactive systems	IS	Computer	3	84	High
Requirement engineering	RE	Computer	3	36	Medium
Software design and construction	SDC	Computer	3	50	Medium
Introduction to computer science	ICS1	Electrical engineering	1	100	Low
Introduction to computer science	ICS2	Electronic engineering	1	198	High
Introduction to computer science	ICS3	Civil engineering	1	85	Low
Introduction to computer science	ICS4	Mining engineering	1	77	Low

Table 2 List of groups by Moodle usage

No	Group	No. of sub-jects
1	High	5
2	Medium	8
3	Low	3

Data and preprocessing

We have used the log data of 1840 Cordoba University students from 16 different courses taught by the Computer Science Department. Table 1 summarises these courses. For each course, it shows the subject or name of the course (Subject), our own identification Code, name of the Degree, Year in the degree/curriculum, number of students (#Users), and the level of Moodle Usage (Low, Medium or High). To accomplish the ethical and privacy issues about using these data, we have used informed consent with all the instructors and we have also anonymized all information about students (Pardo & Siemens, 2014).

We divided or grouped our 16 different courses (see Table 1) into three usage levels of Moodle activities in courses (see Table 2). Moodle provides us resources (text and web page, link to files or websites, and label) and different types of activities (assignments, chat, choice, database, forum, glossary, lesson, quiz, survey, wiki, workshop, etc.). We have defined three levels of usage by the number of activities used in the course:

- *Low level* The course only has one or no activity.
- *Medium level* The course has two different types of activities.
- *High level* The course has three or more different types of activities.

Moodle provides a wide range of activities such as Assignments, Databases, Chats, Choice, Questionnaires, Quiz, Surveys, Forums, Glossaries, Lessons, SCORM packages, Workshops, Wikis, etc.). The most frequent activities in our courses are Assignments, Forums, and Quizzes. So, normally low-level courses only use one of these three activities, medium level two of them, and high level three or more activities. Table 2 shows the number of courses in each group grouped by usage level.

We also propose our ontology for defining 5 high-level attributes starting from 58 low-level attributes or actions provided by Moodle logs (see Table 3).

As depicted in Table 3, our ontology generalizes the 58 raw/low-level events provided by the Moodle logs into only five attributes or high-level categories. The first category references all the actions about consulting resources (LEARNING/READING/VIEWING), the second groups the students' communication events (COMMUNICATING), the third deals with the students' work (WORKING/DOING), the fourth is about students' evaluation (EXAMINING/EVALUATING) and the last is about the students' general ENGAGEMENT in the course. The first four attributes of our ontology are a number (from 0 to 100) that is the percentage of events of each type that each student has done in Moodle. The last attribute is the most general and is also a number (between 0 and 100) obtained from the total number of interactions/events and the number of days connected.

Finally, we have created two different datasets or data files: one with the original previously-described numerical data, and the other discretizing the attributes in two labels (HIGH and LOW) by using the equal width discretization method.

In both cases, we added a new attribute or class to predict at the end of our 5 attributes. This class is the final mark obtained by the students in the course, which is the value to predict in a classification task. The final mark (value between 0 and 10) has been discretized into two values or labels: FAIL if the student's final mark is lower than 5 or PASS if the students' final mark is higher than 5.

Methodology for experimentation

The methodology used in our experimentation consisted of these steps (see Fig. 1):

- Firstly, we downloaded and preprocessed the Moodle log in order to obtain both the numerical and discretized datasets for each course. We used a specific Java tool that we developed for doing this specific transformation task (López-Zambrano et al., 2020).
- Secondly, we executed the well-known J48 classification algorithm provided by the WEKA data mining environment for each one of the previous numerical and

Table 3 Ontology and Moodle low-level actions of each category

Learning/reading/viewing	Communicating	Working/doing	Evaluating/examining	Engagement
Blog view	Forum add discussion	Assignment upload	Hotpot submit	Number of total interactions
Book view all	Forum add post	Assignment view	Hotpot view	
Course enrol	Forum search	Assignment view all	Hotpot view all	Number of days connected
Course recent	Forum subscribe	Assignment view submission	Questionnaire submit	
Course user report	Forum subscribe all	Choice choose	Questionnaire update	
Course view	Forum update	Choice choose again	Questionnaire view	
Folder view	Forum update post	Choice view	Questionnaire view all	
Folder view all	Forum user report	Choice view all	Quiz attempt	
Imsep view all	Forum view discussion	Teamwork update	Quiz close attempt	
Page view	Forum view forum	Teamwork view	Quiz continue attempt	
Page view all	Forum view forums	Teamwork view all	Quiz continue attempt	
Resource view	Wiki edit		Quiz preview	
Resource view all	Wiki update		Quiz review	
Url view	Wiki view		Quiz view	
Url view all	Wiki view all		Quiz view all	

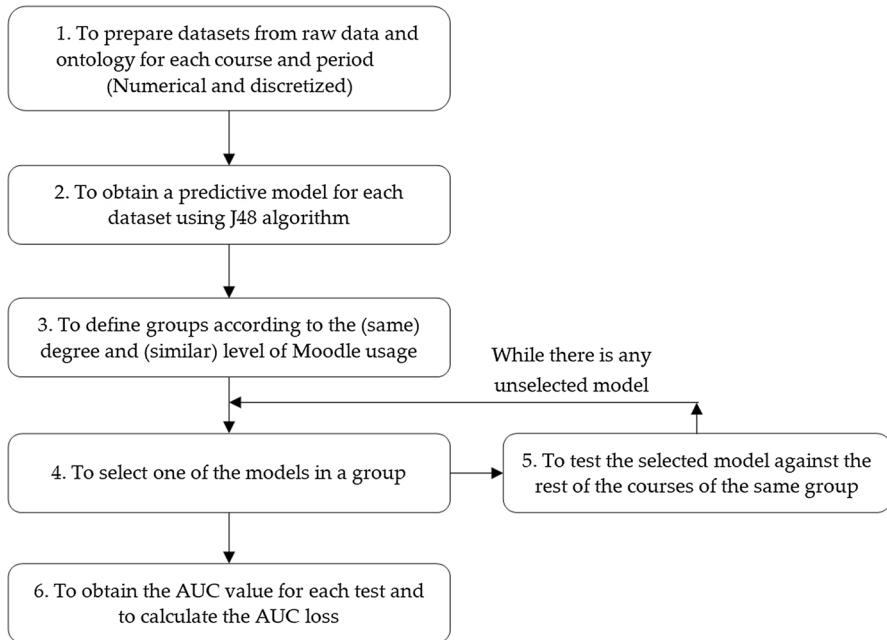


Fig. 1 Methodology used in our experimentation

categorical datasets of 16 subjects or different courses. In this step, we obtained one prediction model for each course.

- Then, we grouped our 16 subjects/courses into 3 groups depending on their level of usage of Moodle activities (see Table 1).
- Next, we repeated the next two actions. We selected each prediction model obtained in one course one by one and we applied it to testing the datasets of all the other courses in the same group. We repeated this process with all the models and with all the datasets for each group.
- Finally, we obtained the values of the two evaluation metrics that we used (the area under the ROC Curve and AUC loss) when applying the prediction model for one course/subject over the other datasets in the same group. And we compared the results obtained when using the original raw low-level

Results

The results of these three groups are set out below (summarised in Table 3). Two experiments were conducted for each group, applying the J48 algorithm with balanced numerical and discretized datasets. These experiments consisted of having a first set of experiments for which high-level datasets were constructed (ontology) and a second experiment with datasets built with low-level attributes.

For each experiment (within the same group), we conducted an analysis of the best AUC obtained and the lowest error rate, or loss of portability, of the model.

Thus, the results consist of two tables. At the top, a matrix is shown with the results of the AUC metric, obtained from the list of the general model for each subject (rows), compared to the average AUC for the individual datasets from each period for a subject (columns). The values of the main diagonal represent the testing of the general model for subjects over their own datasets, where this value is the reference AUC value (highest value), with regard to the AUCs from the other subjects. The second matrix (bottom) displays the difference between the highest AUC (reference) by row, with regard to each individual AUC. These values tell us how much precision is lost in the AUC when this model is tested with other subjects (portability), aiming to highlight the lowest values, as they indicate the lowest error rate or loss in the process of model portability or transferability.

Group of courses with high-level usage

For the high-level group, we can see in Table 4 that of the two tests, the best general results (averages) are in the datasets with ontology, revealing that the AUC average for numerical datasets is 0.62 and the average for discretized datasets is 0.61, higher than their equivalents in the tests without ontology. While there is only a small difference, the loss rate or difference in transferability does denote a greater difference, and within the same test group, the difference between numerical and discretized datasets is highly significant, where the tests with discretized data are much better.

If we focus on the tests with the best results, we can see that the best value for the AUC metric (0.675) is in the ICS2 subject obtained with discretized data and tested with the subject HCI. This is not concordant with the general average of the AUCs, whose highest value is for the numerical sets (0.62), with a tiny difference of one one-hundredth. However, it is concordant with the fact that the best rate of loss or difference is with the discretized data (0.10). We can also see that the generalized model obtained with the aforesaid subject (ICS2) has very good results, which is proven in the general averages (row) in both tests (with and without ontology).

With regard to the model that obtained the best average in the precision loss rate, we can see in Fig. 2, the decision tree, defining the attribute COMMUNICATING (from the five ontology attributes—Table 1) as the attribute with the highest increase in information, which would define the prediction for a student passing the course.

Group of courses with medium-level usage

For the medium-level group, we can see in Table 5 that of the two tests, the best general results (averages) are in the dataset tests with ontology, revealing that the AUC average for numerical datasets is 0.60 and the average for discretized datasets is 0.59, higher than their equivalents in the tests without ontology. There is a small difference, although the loss rate or difference in transferability does denote a greater difference, and within the same test group, the difference between numerical and discretized datasets is highly significant, where the tests with discretized data are much better.

Table 4 AUC results and loss of transferability (difference) with J48—high-level group

Course	With ontology											
	HCI	IS	ICS2	IP2	PM2	Avg	HCI	IS	ICS2	IP2	PM2	Avg
	<i>AUC (numerical datasets)</i>											
HCI	0.890	0.511	0.592	0.535	0.528	0.61	0.710	0.672	0.608	0.614	0.624	0.65
IS	0.488	0.886	0.498	0.555	0.629	0.61	0.509	0.672	0.512	0.576	0.629	0.58
ICS2	0.602	0.600	0.799	0.639	0.661	0.66	0.675	0.633	0.717	0.632	0.662	0.66
IP2	0.483	0.484	0.589	0.849	0.550	0.59	0.536	0.651	0.512	0.704	0.630	0.61
PM2	0.501	0.591	0.483	0.544	0.909	0.61	0.501	0.560	0.562	0.562	0.666	0.57
	Avg mean 0.62											
	<i>AUC loss (numerical datasets)</i>											
HCI	–	0.379	0.298	0.355	0.362	0.35	–	0.038	0.102	0.095	0.085	0.08
IS	0.398	–	0.388	0.331	0.257	0.34	0.163	–	0.159	0.096	0.043	0.12
ICS2	0.197	0.199	–	0.160	0.138	0.17	0.042	0.084	–	0.085	0.055	0.07
IP2	0.366	0.364	0.260	–	0.298	0.32	0.169	0.053	0.192	–	0.074	0.12
PM2	0.408	0.318	0.426	0.364	–	0.38	0.166	0.107	0.104	0.104	–	0.12
	Avg mean 0.31											
	<i>AUC loss (discretized datasets)</i>											
HCI	–	–	–	–	–	–	–	–	–	–	–	–
IS	–	–	–	–	–	–	–	–	–	–	–	–
ICS2	–	–	–	–	–	–	–	–	–	–	–	–
IP2	–	–	–	–	–	–	–	–	–	–	–	–
PM2	–	–	–	–	–	–	–	–	–	–	–	–
	Avg mean 0.10											
Course	Without ontology											
	HCI	IS	ICS2	IP2	PM2	Avg	HCI	IS	ICS2	IP2	PM2	Avg
	<i>AUC (numerical datasets)</i>											
HCI	0.943	0.510	0.522	0.538	0.524	0.61	0.769	0.621	0.569	0.417	0.570	0.59
IS	0.485	0.927	0.494	0.470	0.606	0.60	0.479	0.816	0.577	0.555	0.656	0.62
ICS2	0.514	0.590	0.783	0.500	0.569	0.59	0.503	0.558	0.619	0.485	0.516	0.54
IP2	0.484	0.420	0.472	0.862	0.490	0.55	0.519	0.576	0.535	0.761	0.491	0.58
PM2	0.514	0.489	0.530	0.618	0.899	0.61	0.574	0.488	0.522	0.592	0.793	0.59
	Avg mean 0.59											
	<i>AUC (discretized datasets)</i>											
HCI	–	–	–	–	–	–	–	–	–	–	–	–
IS	–	–	–	–	–	–	–	–	–	–	–	–
ICS2	–	–	–	–	–	–	–	–	–	–	–	–
IP2	–	–	–	–	–	–	–	–	–	–	–	–
PM2	–	–	–	–	–	–	–	–	–	–	–	–
	Avg mean 0.58											

Table 4 (continued)

Course	Without ontology											
	HCI	IS	ICS2	IP2	PM2	Avg	HCI	IS	ICS2	IP2	PM2	Avg
	<i>AUC loss (numerical datasets)</i>					<i>AUC loss (discretized datasets)</i>						
HCI	-	0.432	0.421	0.404	0.418	0.42	-	0.148	0.201	0.352	0.200	0.23
IS	0.442	-	0.433	0.457	0.321	0.41	0.337	-	0.238	0.260	0.160	0.25
ICS2	0.270	0.193	-	0.283	0.215	0.24	0.116	0.061	-	0.134	0.103	0.10
IP2	0.378	0.441	0.390	-	0.371	0.40	0.242	0.184	0.225	-	0.269	0.23
PM2	0.385	0.410	0.369	0.281	-	0.36	0.219	0.305	0.271	0.200	-	0.25
					Avg mean	0.37					Avg mean	0.21

Fig. 2 The best model for the high-level group with discretized dataset—subject ICS2

```
J48 pruned tree
-----
COMMUNICATING = LOW: Fail
COMMUNICATING = HIGH: Pass

Number of Leaves:      2
Size of the tree:      3
```

If we focus on the tests with the best results, we see that the best value for the AUC metric (0.718) is for the subject SDC, obtained with numerical data and tested with the subject RE, which is concordant with the general AUC average, whose highest value is in the numerical tests (0.60), with a small difference of one one-hundredth. However, it is not concordant with the fact that the best loss or difference rate is for discretized data (0.18). We also see that in the generalized model within the tests with discretized data, the subject PM1 has a good result in the general average for the loss rate (row) in the tests without ontology, although in the tests with ontology (employing a generalized model of high-level attributes), there are also good results in the IP1 and INS subjects, which share the same value of 0.13, six one-hundredths more, but still within the ideal value for good transferability of the model.

With regard to the subjects with the best average loss rate, Fig. 3 shows that the decision tree defines the attributes LEARNING/READING/VIEWING and COMMUNICATING (from the five ontology results—Table 1) as the attributes with the greatest gain in information, defining that if there is a high level of LEARNING/READING/VIEWING, the student will pass or, conversely, if it is low, but with a high level of interaction in COMMUNICATING, the student will also pass.

Concerning the decision tree shown in Fig. 4, it defines the attributes LEARNING/READING/VIEWING, COMMUNICATING, WORKING/DOING and EVALUATING/EXAMINING (from the five ontology attributes—Table 1) as the attributes with the greatest increase in information, once again defining that if there is a high level of LEARNING/READING/VIEWING, the student will pass or, conversely, if it is low, but with a high level of interaction in COMMUNICATING, the student will also pass. If the COMMUNICATING level is low, but the level of WORKING/DOING is high, then the student would pass, but if it is not high, then the student will only pass if the EVALUATING/EXAMINING level is high.

Group of courses with low-level usage

For the low-level group, we can see in Table 6 that of the two tests, the best general results (averages) are in the dataset tests with ontology, revealing that the AUC averages for numerical datasets is 0.63 and the average for discretized datasets is 0.61, higher than their equivalents in the tests without ontology. There is a small difference, although the loss rate or difference in transferability does denote a greater difference, and within the same test group, the difference between numerical and discretized datasets is highly significant, where the tests with discretized data are much better.

Table 5 AUC results and loss of transferability (difference) with J48—medium-level group

Course	With ontology								
	IP1	PM1	DB	SDC	PCT	RE	SE	InS	Avg
<i>AUC (numerical datasets)</i>									
IP1	0.835	0.567	0.589	0.552	0.508	0.589	0.620	0.582	0.61
PM1	0.519	0.821	0.540	0.520	0.530	0.510	0.550	0.567	0.57
DB	0.670	0.623	0.980	0.590	0.571	0.566	0.521	0.640	0.65
SDC	0.502	0.596	0.516	0.788	0.469	0.718	0.549	0.504	0.58
PCT	0.633	0.621	0.611	0.610	0.911	0.641	0.572	0.670	0.66
RE	0.494	0.519	0.497	0.643	0.476	0.869	0.527	0.512	0.57
SE	0.540	0.510	0.520	0.560	0.530	0.511	0.962	0.523	0.58
InS	0.608	0.580	0.591	0.563	0.508	0.560	0.562	0.815	0.60
								Avg mean	0.60
<i>AUC loss (numerical datasets)</i>									
IP1	–	0.267	0.246	0.283	0.327	0.246	0.215	0.253	0.26
PM1	0.302	–	0.281	0.301	0.291	0.311	0.271	0.254	0.29
DB	0.310	0.357	–	0.390	0.409	0.414	0.459	0.340	0.38
SDC	0.286	0.192	0.272	–	0.319	0.070	0.239	0.284	0.24
PCT	0.278	0.290	0.300	0.301	–	0.270	0.339	0.241	0.29
RE	0.375	0.350	0.372	0.226	0.393	–	0.342	0.357	0.34
SE	0.422	0.452	0.442	0.402	0.432	0.451	–	0.439	0.43
InS	0.207	0.235	0.224	0.252	0.307	0.255	0.253	–	0.25
								Avg mean	0.31
<i>AUC (discretized datasets)</i>									
IP1	0.772	0.637	0.621	0.601	0.688	0.643	0.643	0.652	0.66
PM1	0.634	0.763	0.532	0.604	0.562	0.510	0.521	0.602	0.59
DB	0.612	0.583	0.775	0.555	0.616	0.567	0.543	0.551	0.60
SDC	0.474	0.562	0.628	0.696	0.505	0.590	0.480	0.551	0.56
PCT	0.592	0.564	0.577	0.582	0.812	0.567	0.581	0.582	0.61
RE	0.589	0.591	0.520	0.583	0.572	0.801	0.563	0.571	0.60
SE	0.527	0.562	0.550	0.588	0.504	0.614	0.694	0.548	0.57
InS	0.648	0.635	0.549	0.640	0.471	0.369	0.529	0.677	0.56
								Avg mean	0.59
<i>AUC loss (discretized datasets)</i>									
IP1	–	0.135	0.151	0.172	0.084	0.129	0.129	0.120	0.13
PM1	0.129	–	0.231	0.159	0.201	0.253	0.242	0.162	0.20
DB	0.163	0.192	–	0.220	0.159	0.208	0.232	0.224	0.20
SDC	0.222	0.134	0.068	–	0.191	0.107	0.216	0.145	0.15
PCT	0.220	0.248	0.235	0.230	–	0.245	0.231	0.230	0.23
RE	0.212	0.210	0.281	0.218	0.229	–	0.238	0.230	0.23
SE	0.167	0.132	0.144	0.107	0.190	0.080	–	0.146	0.14
InS	0.029	0.042	0.128	0.038	0.206	0.309	0.148	–	0.13
								Avg mean	0.18

Table 5 (continued)

Course	Without ontology								
	IP1	PM1	DB	SDC	PCT	RE	SE	InS	Avg
<i>AUC (numerical datasets)</i>									
IP1	0.938	0.588	0.542	0.545	0.610	0.493	0.579	0.523	0.60
PM1	0.496	0.689	0.589	0.478	0.567	0.624	0.484	0.486	0.55
DB	0.495	0.491	0.976	0.535	0.457	0.670	0.581	0.517	0.59
SDC	0.492	0.518	0.467	0.809	0.504	0.558	0.496	0.456	0.54
PCT	0.459	0.496	0.337	0.585	0.891	0.612	0.382	0.492	0.53
RE	0.439	0.524	0.329	0.553	0.579	0.956	0.473	0.577	0.55
SE	0.526	0.581	0.611	0.559	0.486	0.614	0.964	0.494	0.60
InS	0.484	0.495	0.671	0.583	0.486	0.610	0.533	0.704	0.57
								Avg mean	0.57
<i>AUC loss (numerical datasets)</i>									
IP1	–	0.350	0.396	0.393	0.328	0.446	0.359	0.415	0.38
PM1	0.193	–	0.100	0.211	0.122	0.065	0.205	0.203	0.16
DB	0.481	0.485	–	0.441	0.519	0.307	0.395	0.459	0.44
SDC	0.317	0.291	0.342	–	0.305	0.252	0.313	0.353	0.31
PCT	0.432	0.395	0.554	0.306	–	0.279	0.509	0.399	0.41
RE	0.517	0.432	0.627	0.403	0.377	–	0.483	0.379	0.46
SE	0.438	0.383	0.353	0.405	0.478	0.351	–	0.470	0.41
InS	0.221	0.209	0.033	0.121	0.218	0.094	0.171	–	0.15
								Avg mean	0.34
<i>AUC (discretized datasets)</i>									
IP1	0.811	0.441	0.496	0.535	0.500	0.500	0.414	0.510	0.53
PM1	0.476	0.585	0.458	0.550	0.515	0.564	0.512	0.559	0.53
DB	0.551	0.500	0.652	0.551	0.476	0.500	0.510	0.499	0.53
SDC	0.532	0.593	0.430	0.924	0.531	0.610	0.484	0.622	0.59
PCT	0.494	0.500	0.447	0.567	0.712	0.553	0.470	0.551	0.54
RE	0.568	0.543	0.529	0.614	0.508	0.756	0.545	0.569	0.58
SE	0.487	0.500	0.500	0.375	0.473	0.431	0.718	0.451	0.49
InS	0.526	0.500	0.429	0.625	0.528	0.454	0.500	0.761	0.54
								Avg mean	0.54
<i>AUC loss (discretized datasets)</i>									
IP1	–	0.370	0.315	0.277	0.311	0.311	0.397	0.301	0.33
PM1	0.108	–	0.127	0.035	0.070	0.021	0.073	0.025	0.07
DB	0.101	0.152	–	0.101	0.176	0.152	0.142	0.153	0.14
SDC	0.392	0.331	0.494	–	0.393	0.314	0.440	0.302	0.38
PCT	0.218	0.212	0.265	0.145	–	0.159	0.242	0.161	0.20
RE	0.188	0.213	0.227	0.142	0.248	–	0.211	0.187	0.20
SE	0.231	0.218	0.218	0.343	0.245	0.287	–	0.267	0.26
InS	0.235	0.261	0.332	0.136	0.233	0.307	0.261	–	0.25
								Avg mean	0.23

Fig. 3 Best model for the medium-level group with discretized dataset—subject IP1

```
J48 pruned tree
-----
LEARNING/READING/VIEWING = LOW
| COMMUNICATING = LOW: Fail
| COMMUNICATING = HIGH: Pass
LEARNING/READING/VIEWING = HIGH: Pass

Number of Leaves:      3
Size of the tree:     5
```

Fig. 4 Best model for the medium-level group with discretized dataset—subject InS

```
J48 pruned tree
-----
LEARNING/READING/VIEWING = LOW
| COMMUNICATING = LOW
| | WORKING/DOING = LOW
| | | EVALUATING/EXAMINING = LOW: Fail
| | | EVALUATING/EXAMINING = HIGH: Pass
| | WORKING/DOING = HIGH: Pass
| COMMUNICATING = HIGH: Pass
LEARNING/READING/VIEWING = HIGH: Pass

Number of Leaves:      5
Size of the tree:     9
```

If we now focus on the tests with the best results, we can see that the best value for the AUC metric (0.683) is in the ICS1 subject obtained with discretized data and tested with the subject ICS4. This is not concordant with the general average of the AUCs whose highest value is for the numerical sets (0.63), with a small difference of two one-hundredths. However, it is concordant with the fact that the best rate of loss or difference is with the discretized data (0.13). We can also observe that the generalized model obtained with the aforesaid subject (ICS1) has very good results, which is proven in the general averages (row) for the matrix of discretized data with ontology. The value is 0.07, which is below the ideal for determining a good transfer of the model, in this case for a general model with high-level attributes.

With regard to the subject with the best average loss rate, we can see in Fig. 5, the decision tree, defining the attribute COMMUNICATING (from the five attributes of ontology—Table 1) as the attribute with the highest increase in information, which would define the predictability for a student passing the course.

Conclusions

This paper aims to improve the portability or transferability of predictive models of students' performance by using an ontology that uses a taxonomy of actions on students' interactions with the Moodle learning management system. We compare the results of this new proposed approach against our previous results when we used low-level raw attributes directly obtained from Moodle logs. The results obtained

Table 6 AUC results and loss of transferability (difference) with J48—low-level group

Course	With ontology							
	ICS1	ICS3	ICS4	Avg	ICS1	ICS3	ICS4	Avg
	<i>AUC (numerical datasets)</i>				<i>AUC (discretized datasets)</i>			
ICS1	0.860	0.592	0.500	0.65	0.722	0.615	0.683	0.67
ICS3	0.506	0.820	0.560	0.63	0.512	0.750	0.565	0.61
ICS4	0.510	0.531	0.832	0.62	0.500	0.500	0.600	0.53
			Avg mean	0.63			Avg mean	0.61
	<i>AUC loss (numerical datasets)</i>				<i>AUC loss (discretized datasets)</i>			
ICS1	–	0.268	0.360	0.31	–	0.107	0.039	0.0
ICS3	0.314	–	0.260	0.29	0.239	–	0.186	0.21
ICS4	0.322	0.301	–	0.31	0.100	0.100	–	0.10
			Avg mean	0.30			Avg mean	0.13
Course	Without ontology							
	ICS1	ICS3	ICS4	Avg	ICS1	ICS3	ICS4	Avg
	<i>AUC (numerical datasets)</i>				<i>AUC (discretized datasets)</i>			
ICS1	0.917	0.491	0.404	0.60	0.761	0.470	0.591	0.61
ICS3	0.554	0.938	0.527	0.67	0.375	0.707	0.502	0.53
ICS4	0.414	0.495	0.771	0.56	0.410	0.460	0.682	0.52
			Avg mean	0.61			Avg mean	0.55
	<i>AUC loss (numerical datasets)</i>				<i>AUC loss (discretized datasets)</i>			
ICS1	–	0.426	0.513	0.47	–	0.291	0.170	0.23
ICS3	0.384	–	0.411	0.40	0.333	–	0.205	0.27
ICS4	0.357	0.277	–	0.32	0.273	0.222	–	0.25
			Avg mean	0.39			Avg mean	0.25

Fig. 5 Best model for the low-level group with discretized dataset—subject ICS1

```

J48 pruned tree
-----
COMMUNICATING = LOW: Fail
COMMUNICATING = HIGH: Pass

Number of Leaves:      2
Size of the tree:      3

```

show that the use of the proposed ontology significantly improves the portability of the models in terms of their predictive accuracy. So, the answer to our initial research question is yes, the ontological models obtained in one source course can be applied to other different target courses with similar usage levels without losing prediction accuracy.

One of the limitations of this work is the specific attributes/variables used in our proposed ontology.

For example, it is also important to discuss if the “number of total interactions” are truly showing engagement when learning using LMS. The number of actions

includes the behavior of supposed relevant activity in the LMS and were assuming that all of these actions could indicate that the student is properly involved in his learning process. As traditionally happens with study time, however, this variable by itself is very tricky. It may seem that the more time those students spend studying, the better grades they should receive, but it is not that simple; it mainly depends on the quality of the study time, and something similar could be occurring with the relevant actions; more activity in the LMS does not assure better results (Cerezo et al., 2016).

Regarding the application of the results obtained in this work and the potential for using them within other domains; it is important to notice that currently there is an increasing interest in the generalization and portability of prediction models and specifically with Moodle LMS (Monllao-Olive et al., 2019). In this line, our proposal can be applied not only to Learning Management Systems as Moodle but also to other different domains or data sources such as Intelligent Tutoring Systems (ITSs), Massive Online Open Courses (MOOCs), Traditional face-to-face educational environments, Blended Learning and Multimodal Learning environments, and so on.

Finally, as a future study, we are currently working on:

- Using a higher number of courses with much more data/students from different areas/domains, not only engineering and computer science, but also fields such as science, biology, medicine, philosophy, and literature, in order to generalize the good results that we obtained in this study.
- Discovering predictive models that can be portable/transferable as soon as possible in the early stages of the course. This means we would not have to wait until the end of the course to have all Moodle usage data available, and the obtained models could be used as general early warning prediction models for different similar courses (Romero & Ventura, 2019).

Acknowledgements This work would not be possible without the funding from the Ministry of Sciences and Innovation I+D+I TIN2017-83445-P.

References

- Al-Yahya, M., George, R., & Alfaries, A. (2015). Ontologies in E-learning: Review of the literature. *International Journal of Software Engineering and Its Applications*, 9(2), 67–84.
- Baker, R. S. (2019). Challenges for the future of educational data mining: The baker learning analytics prizes. *Journal of Educational Data Mining*, 11, 1–17.
- Boyer, S., & Veeramachaneni, K. (2015). Transfer learning for predictive models in massive open online courses. In C. Conati, N. Heernan, A. Mitrovic, & M. Verdejo (Eds.), *Artificial intelligence in education. AIED 2015. Lecture notes in computer science*. (Vol. 9112). Springer.
- Castro, F. A., & Alonso, M. A. (2011). Learning objects and ontologies to perform educational data mining. In *International conference on frontiers in education: Computer science and computer engineering*.
- Cerezo, R., Sánchez-Santillán, M., Paule-Ruiz, M. P., & Núñez, J. C. (2016). Students' LMS interaction patterns and their relationship with achievement: A case study in higher education. *Computers & Education*, 96, 42–54.

- Cerezo, R., Boragin, A., Esteban, M., & Romero, C. (2020). Process mining for self-regulated learning assessment in E-learning. *Journal of Computing in Higher Education*, 32, 74–88.
- Chang, M., D'Aniello, G., Gaeta, M., Orciuoli, F., Sampson, D., & Simonelli, C. (2020). Building ontology-driven tutoring models for intelligent tutoring systems using data mining. *IEEE Access*, 8, 48151–48161.
- Dhuria, S., & Chawla, S. (2014). Ontologies for personalized E-learning in the semantic web. *International Journal of Advanced Engineering and Nano Technology*, 1(4), 13–18.
- Ding, M., Wang, Y., Hemberg, E., & O'Reilly, U.-M. (2019). Transfer learning using representation learning in massive open online courses. In *Proceedings of the 9th international conference on learning analytics & knowledge (LAK19), Empe, AZ, USA, 4–8 March 2019* (pp. 145–154). ACM.
- Dorça, F. A., Carvalho, V. C., Mendes, M. M., Araújo, R. D., Ferreira, H. N., & Cattelan, R. G. (2017). An approach for automatic and dynamic analysis of learning objects repositories through ontologies and data mining techniques for supporting personalized recommendation of content in adaptive and intelligent educational systems. In *IEEE 17th international conference on advanced learning technologies (ICALT)* (pp. 514–516).
- Dougiamas, M., & Taylor, P. (2008). Moodle: Using learning communities to create an open source course management system. In *EdMedia: World conference on educational media and technology* (pp. 171–178). The LearnTechLib.
- El-Rady, A. A. (2020). An ontological model to predict dropout students using machine learning techniques. In *3rd international conference on computer applications & information security (ICCAIS)* (pp. 1–5).
- Fong, S., Ip, A., & Mohammed, S. (2011). *A multi-level biomedical classification model by using aggregation and abstraction techniques*. Biomedical Engineering (Biomed 2011).
- Forehand, M. (2005). Bloom's taxonomy: Original and revised. In M. Orey (Ed.), *Emerging perspectives on learning, teaching, and technology (E-Book)*. Retrieved January 15, 2021, from https://textbookequity.org/Textbooks/Orey_Emergin_Perspectives_Learning.pdf
- Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, 28, 68–84.
- Grivokostopoulou, F., Perikos, I., & Hatzilygeroudis, I. (2014). Utilizing semantic web technologies and data mining techniques to analyze students learning and predict final performance. In *2014 IEEE international conference on teaching, assessment and learning for engineering (TALE)* (pp. 488–494).
- Hunt, X. J., Kabul, I. K., & Silva, J. (2017). Transfer learning for education data. In *Proceedings of the ACM SIGKDD conference*.
- López-Zambrano, J., Lara, J. A., & Romero, C. (2020). Towards portability of models for predicting students' final performance in university courses starting from Moodle Logs. *Applied Science*, 10(354), 1–23.
- Marinho, T., Costa, E. B., Dermeval, D., Ferreira, R., Braz, L. M., Bittencourt, I. I., & Luna, H. P. L. (2010). An ontology-based software framework to provide educational data mining. In *Proceedings of the 2010 ACM symposium on applied computing (SAC '10). Association for computing machinery* (pp. 1433–1437).
- Monllao-Olive, D. M., Huynh, D. Q., Reynolds, M., Dougiamas, M., & Wiese, D. (2019). A quest for a one-size-fits-all neural network: Early prediction of students at risk in online courses. *IEEE Transactions on Learning Technologies*, 12(2), 171–183.
- Montenegro-Marin, C. E., Cueva-Lovelley, J. M., Sanjuan, O., & Nuñez-Valdez, E. R. (2011). Towards an ontology to describe the taxonomy of common modules in learning management systems. In *International journal of interactive multimedia and artificial intelligence, 1, special issue on computer science and software engineering* (pp. 47–53).
- Nouira, A., Cheniti-Belcadhi, L., & Braham, R. (2019). An ontology-based framework of assessment analytics for massive learning. *Computer Applications on Engineering Education*, 27, 1343–1360.
- Octaviani, D., Pranolo, A., & Othman, S. (2015). RDB2Onto: An approach for creating semantic meta-data from relational educational data. In *International conference on science and information technology (ICSITech)* (pp. 137–140).
- Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, 45(3), 438–450.
- Rollins, M. (2010). *Moodle and Bloom's taxonomy*. Lulu Enterprise.

- Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, 51, 368–384.
- Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3, 12–27.
- Romero, C., & Ventura, S. (2019). Guest editorial: Special issue on early prediction and supporting of learning performance. *IEEE Transactions on Learning Technologies*, 12(2), 145–147.
- Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. In *Wiley interdisciplinary reviews: Data mining and knowledge discovery*. Wiley.
- Tang, A., & Fong, S. (2010). A taxonomy-based classification model by using abstraction and aggregation. In *2010 6th international conference on advanced information management and service (IMS)* (pp. 448–454).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Javier López-Zambrano is Assistant Professor at Escuela Superior Politécnica Agropecuaria de Manabí (ESPAM MFL), Ecuador. He is currently Ph.D. student at University of Córdoba, UCO, Spain, where he is conducting research on early prediction of students' academic performance by using data mining techniques. His research interests in computer science include data mining, knowledge discovery in databases, and e-learning.

Juan A. Lara is Associate Professor and Research Scientist at Madrid Open University, UDIMA, Spain. He is currently member of Department of Computer. He holds a Ph.D. in Computer Science and two Post Graduate Masters in Information Technologies and Emerging Technologies to Develop Complex Software Systems from Technical University of Madrid, Spain. He is author of more than 30 papers published in international impact journals. He has served as guest editor in a number of special issues of high impact factor journals and organized one international conference. His research interests in computer science include data mining, knowledge discovery in databases, data fusion, artificial intelligence and e-learning.

Cristóbal Romero is Full Professor at the University of Córdoba in Spain and member of the KDIS (Knowledge Discovery and Intelligent Systems) research group. He has published more than 150 papers in books, journals and conferences, 50 of which have been published in Thomson-Reuters Impact Factor (IF) journals and some of them are important EDM (Educational Data Mining) surveys/reviews. He was also the co-editor of several special issues and two books regarding EDM specific topics. He was a founding officers of the international EDM society and he has served in the program committee of a great number of international conferences about education, personalization, data mining, AI and he has organized as Program Chair and Conference Chair several editions of EDM conference. His main research interests are the application of data mining/learning analytics and artificial intelligence techniques to educational data/environment/domain.