Ros, S., Lázaro, J. C., Robles-Gómez, A., Caminero, A. C., Tobarra, L., & Pastor, R. (2017). Analyzing Content Structure and Moodle Milestone to Classify Student Learning Behavior in a Basic Desktop Tools Course. In J. M. Dodero, M. S. Ibarra Sáiz, & I. Ruiz Rube (Eds.), Fifth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'17) (Cádiz, Spain, October 18-20, 2017) (pp. Article 42). New York, NY, USA: ACM. doi:10.1145/3144826.3145392

Analyzing Content Structure and Moodle Milestone to Classify Student Learning Behavior in a Basic Desktop Tools Course

Salvador Ros
Universidad Nacional de Educación a Distancia (UNED)
Madrid, Spain
sros@scc.uned.es

Juan Carlos Lázaro
Universidad Nacional de Educación a Distancia (UNED)
Madrid, Spain
jclo@scc.uned.es

Antonio Robles-Gómez
Universidad Nacional de Educación a Distancia (UNED)
Madrid, Spain
arobles@scc.uned.es

Agustín C. Caminero
Universidad Nacional de Educación a Distancia (UNED)
Madrid, Spain
accaminero@scc.uned.es

Llanos Tobarra
Universidad Nacional de Educación a Distancia (UNED)
Madrid, Spain
llanos@scc.uned.es

Rafael Pastor
Universidad Nacional de Educación a Distancia (UNED)
Madrid, Spain
rpastor@scc.uned.es

ABSTRACT

This paper analyzes the content structure and Moodle milestone to classify the students' learning behavior for a basic desktop-tools on-line virtual course. The data collection phase is completed for a Learning Analytics (LA) process as a first step; by using the generated interactions among students, and with learning resources, assessments, and so on. A first exploratory data analysis study is also done with the extracted indicators (or features) of all interactions to classify them in five traits. A multidimensional parameter reduction has been implemented based on Principal Component Analysis (PCA), an example of it is also given.

CCS CONCEPTS

• **Social and professional topics → Computer science education**; • **Computing methodologies → Classification and regression trees**; • **Applied computing → E-learning**;

KEYWORDS

Learning Analytics (LA), Principal Component Analysis (PCA), Exploratory Data Analysis (EDA), Moodle, Indicators

1 INTRODUCTION

Nowadays, with the evolution of technology in the field of virtual learning, a new learning model, named Learning Analytics (LA) [9], has emerged. The entire LA process is composed by four different phases, each of them acting as the entry point for the next phase. These phases are data capture, extraction of relevant information, processing it for further analysis, and taking a set of actions. These actions will be taken into account for the next refined LA cycle, during the data capture phase.

On the other hand, Learning Management Systems (LMS) have nowadays become an excellent platform to teach on distance. In this sense, a big amount of data is generated when students interact with other students and content resources provided by the platform. All these data can be captured from the existing platforms to be used for a LA process, as this work does. There are available several LMSs for the academic community, both proprietary and open-source, such as Moodle, Sakai, Blackboard, and so on.

In this work Moodle is employed. This is an open source, a well-known platform, which has been used by a great number of learning institutions and Universities in recent years. This platform is equipped with a basic analytic module and a log filesystem, providing with great pieces of information about how each student is working on the platform. Part of this information is easily accessible if the content structure is designed for this purpose before starting a virtual course.

Analyzing the different parameters that are recorded by the log system and the content structure in Moodle for an online course, it is possible to classify those in different learning traits. These traits allow us to study the students' learning behavior and classify them in order to understand how they are learning, how to improve the content structure and students' grades. As a consequence, to help students to minimize bad practices, as well as detecting drops out or the loneliness of a distance student.

In order to deal with all the collected information, a multidimensional parameter reduction has been implemented based on PCA (Principal Component Analysis) [3]. Therefore, only the most relevant parameters of each kind of trait have been selected in order to classify students. In this paper, the exploratory analysis of data (EDA) is performed by classifying the collected data and filtering the collected characteristics with PCA. In addition to this, the first results and conclusions obtained are described by using a particular trait.

The structure of this paper is as follows. Section II describes the state of the art for PCA and the data collection phase focusing on Moodle. After that, Section III provides with the analysis of the extracted indicators and the PCA reduction. Finally, Section IV discusses our conclusions and suggests guidelines for future work.

2 STATE OF THE ART

2.1 Principal Component Analysis (PCA)

In general, LA consists on measuring, collecting, analyzing, and reporting data about students and their general context (such as the belonging course, faculty, and even institution) [7]. This way, the learning process can be studied deeper and, also, optimized according to the context where the learning process is applied. During the data collection phase, the captured data must be filtered in order to produce and keep the most relevant information, and so that the subsequent processing of information is not computationally so complex.

The main objective of this work, once the data capture is performed, is to filter and organize the data obtained from a virtual course, where the learning/teaching process is on distance. Therefore, a later processing of information and classification will be more efficient and the made decisions about the virtual course will be easier. To achieve this, it is necessary to reduce the dimensionality of the captured data and to extract relevant characteristics (or indicators) from each subset of correlated learning indicators. This way, the latter processing of information will be computationally affordable.

One of the more suitable techniques for the reduction of dimensionality and extraction of characteristics of our dataset is PCA (Principal Component Analysis). The analysis of principal components is one of the most popular techniques for the reduction of characteristics (or indicators), remaining with the most relevant ones. This reduction is usually reduced to 1 or 2 features from each subset of characteristics. The algorithm used in this work is the traditional PCA, with linear transformation, commonly known as "attribute subset selection" [2]. There are in the literature other PCA algorithms that are more sophisticated in accuracy, but with the penalization of a greater computational complexity.

It would have also been possible to use another approach to generate a new reduced set of characteristics, in which each of them would represent several original features. For example, for a set of subjects from a course (mathematics, physics, history, language...), new features could be created. In particular, they could be sciences and humanities; mathematics, physics, and so on, would belong to science, and history, language, and so on, to humanities. In this paper, this approach has been discarded since our objective is to obtain the most representative characteristics, even if there is information loss. This is not a real problem, in our case, because the extracted characteristics have been grouped into several subsets, the characteristics or indicators of each subset will be highly correlated statistically. If a set of parameters is seven or eight, PCA gathers the two or three most relevant features.

Other techniques that could perform a similar work, such as SVD (Singular Value Decomposition) [3] or NNMF (Non-Negative Matrix Factorization) [8], should be studied in a nearby future to check their precision during the filtering task, and to compare to the current proposal.

2.2 Data Collection in On-line Platforms

Learning Management Systems (LMSs) are on-line platforms, which allow students and faculty to access the learning/teaching resources of a virtual course. At distance education, as the case of UNED (Spanish University for Distance Education; In Spanish, Universidad Nacional de EducaciÃşn a Distancia), the LMS is the main element of interaction among students and faculty. The LMS includes both communication tools (asynchronous forums and chats), as well as evaluation tools, and tracking tools for monitoring the evolution of students, so easing the creation of social knowledge. All these tools try to replace traditional physical interaction. For this reason, it is necessary to study all generated information during the learning process in on-line platforms, thus improving virtual interaction.

All current LMS platforms store a set of parameters/indicators closely related to learning resources, assessment, interaction tools and tracking of students. In this work, each of these parameters will be seen as characteristics collected from a Moodle platform for starting a LA process. Nowadays, some institutions have also started massive courses in MOOCs platforms. In these platforms, some recent studies have been completed about the students' behaviors in virtual courses, from the point of view of massive courses, such as Coursera [5], OpenEdX [6], or OpenMOOC [10]. The last one is the UNED platform for MOOCs courses. These works have taken into account for our proposed work to improve the classification of indicators in traditional LMSs.

In the case of Moodle, which is the LMS employed in this study, the activity recording system captures the data and stores them in the LMS's database. Specifically, all students' interactions in the LMS are stored in a log that improves the scalability of the system, monitors information obtained and, also, it is able to store information on external platforms for later filtering, analysis and decision making [1]. The Moodle platform contains two types of APIs, one for events and another one for registry. The second one includes four plug-ins: log manager, log store, external database log store, and legacy log reader. More details on the internal working of these Moodle APIs and their management and communication can be found in [1].

As described by [4], the information modules that can be extracted from Moodle are related to content interaction (Assignment, Course, Notes, Resource, Upload, Quiz) and interpersonal interaction (Blog, Choice, Forum). This classification is too simple, and it does not take into account the students' traits. In next section, an improved classification is proposed by taken into account traits.

3 EXPLORATORY DATA ANALYSIS

3.1 Experiment Setup

The course selected for this study is a basic desktop tools course where the use of word processor, spreadsheet, and so on (five modules in total), are introduced to students. The course is divided

in different modules, in which all of them have the same minimum structure: multimedia interactive content, basic documentation and assessment. The course has also a general part where all the information needed about the use of the platform, evaluation processes and additional useful information for the student is described. From the collected data in the LA process, an exploratory data analysis (EDA) is performed. As a first task, a total of 70 anonymized parameters/indicators, which can be seen as characteristics, have been recorded and used to define five learning traits.

This classification of features is proposed by taking into account the traditional organization of LMSs and the philosophy of MOOCs, and it is as follows:

(1) Pre-study: the student reads the general information for the course, and knows how to deal with it;
(2) Basic study: the student uses the specific course contents;
(3) Evaluation: the student makes assessments and practices;
(4) Communication: the student uses the communication tools for the study by asking for questions and helping other students and, finally,
(5) Implication: the student spends time regularly in the course.

3.2 Extracted Indicators

The extracted indicators from the virtual course from Moodle are shown in Table 1, and grouped by each trait. As observed, the pre-study trait contains indicators related to number of accesses to surveys, guides about the general course, software, installations, and news. The basic study contains indicators about the access to general contents, multimedia learning, theoretical contents, and specific guides. As for the evaluation, indicators are thought about partial qualifications about each module, the whole module, and the total final grade. With respect to number of students' interactions in the virtual course, several general forums area available, as well as specific contents about a set of contents of modules. A more novel trait is the implication, in which students' actions, intervals, dedication, are recorded, among others.

Table 1: Extracted Indicators Organized by Traits

Trait	Indicators
Pre-study(9 indicators)	Quality survey
	Expectative survey
	Course study guide
	Course user guide
	Course guide
	Software guide
	General instructions
	News
	Videos for software installation
Basic study(16 indicators)	About the course
	Op. Systems multimedia resources (Module 1)
	Word processing multimedia resources
	(Module 2)
	Spreadsheet multimedia resources (Module 3)
	Presentation multimedia resources (Module 4)
	Database multimedia resources (Module 5)
	Study guide of module 1
	Study guide of module 2
	Study guide of module 3
	Study guide of module 4
	Study guide of module 5
	Theoretical contents of module 1
	Theoretical contents of module 2
	Theoretical contents of module 3
	Theoretical contents of module 4
	Theoretical contents of module 5
Evaluation(26 indicators)	Total final qualification
	Final qualification 1 (Module 1)
	Final qualification 2 (Module 2)
	Final qualification 3 (Module 3)
	Final qualification 4 (Module 4)
	Final qualification 5 (Module 5)
	Partial qualification 1a
	Partial qualification 1b
	Partial qualification 1c
	Partial qualification 1d
	Partial qualification 2a
	Partial qualification 2b
	Partial qualification 2c
	Partial qualification 2d
	Partial qualification 3a
	Partial qualification 3b
	Partial qualification 3c
	Partial qualification 3d
	Partial qualification 4a
	Partial qualification 4b
	Partial qualification 4c
	Partial qualification 4d
	Partial qualification 5a
	Partial qualification 5b
	Partial qualification 5c
	Partial qualification 5d
Communication(13 indicators)	Debate
	General forum
	Student forum
	Specific forum 1
	Specific forum 2
	Specific forum 3

Continued on next column

Continued from previous column

Continued from previous column

Trait	Indicators
	Specific forum 4
	Specific forum 5
	Specific forum 6
	Specific forum 7
	Specific forum 8
	Percentage forum
	Messages
Implication(6 indicators)	Actions
	Interval
	Dedication
	Days
	Regularity
	Dedication (by minutes)

Concluded

The above mentioned indicators are a refined classification from the found in the literature, and more specifically in [4]. The set of parameters of each trait are also very detailed.

3.3 Looking for student behaviors traits

LMSs, as Moodle, depending on the content instructional design is really high and then difficult to interpret. For this purpose, a PCA based methodology has been proposed in order to reduce the dimensionality of the problem and as first step to develop an automatic tool for improving teaching learning processes based on learning analytics.

First of all, the parameters are classified based on the previous knowledge that it is possible to collect from the instructional design of the contents. This can understand as previous knowledge constraint useful for a first step in order to detect the most interesting behaviors traits. Despite this first classification the number of parameters for each trait is still great.

The method applied is:

(1) To analyze the correlation matrix of each trait with the goal of extract from the trait the less correlated parameters. This parameters with low correlation among them could be identified as components of new hidden traits.
(2) Apply PCA algorithm to detect the principal components to reduce de dimensionality of the problem.
(3) Based on the results, applied classification algorithms to the data set to detect different states for any trait proposed. The result of this steps is a classification of a students for a concrete trait. This method applies in turn for each trait, (previous detected and the new one result of different discards) builds a trait based profile for each student.

As an example, the method has been applied to pre-study trait. This trait is about how a student prepare herself to begin her course. Parameters grouped under this trait show the next correlation graph (Figure 1).

Table 2: Pre-study Trait Parameters Correlation Matrix

	StudyG	StudyVG	CGuide	SoftG	News	Vtutorial
StudyG	1.00	0.49	0.60	0.34	0.40	0.29
StudyVG	0.49	1.00	0.62	0.40	0.36	0.32
CGuide	0.60	0.62	1.00	0.35	0.39	0.27
SoftG	0.34	0.40	0.35	1.00	0.45	0.62
News	0.40	0.36	0.39	0.45	1.00	0.38
Vtutorial	0.29	0.32	0.27	0.62	0.38	1.00

Table 3: Matrix Component

	PCA1	PCA2	PCA3
StudyG	-0,73	-0,38	-0,08
StudyVG	-0,75	-0,31	0,28
CGuide	-0,76	-0,45	0,08
SoftG	-0,73	0,49	0,15
News	-0,68	0,16	-0,69
Vtutorial	-0,65	0,59	0,23

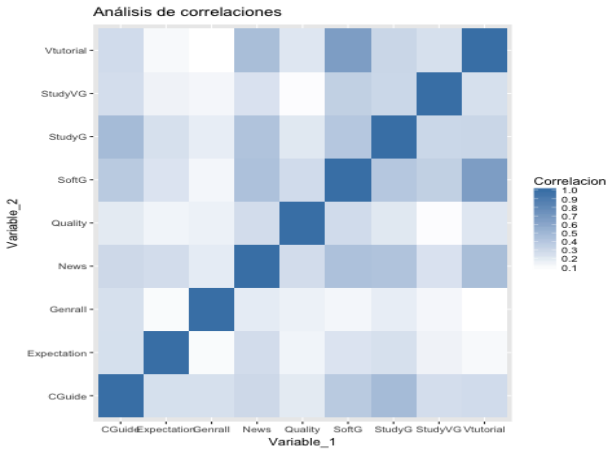

Figure 1: Pre-study Trait Correlation Graph

From this graph it is easy to find out that there are at least three parameters with a low correlation: Quality survey (1), Expectative survey (2) and General information (7). This result is predictable since they are parameters that collect information about the participation of the student in these surveys more than give information about how prepare themselves for the study. In this example, the parameter General information give a very short information about the content of the course. Then it is possible to consider discard them in the process or to be part of new trait. If the three parameters are discarded the result of calculating the new correlation matrix is shown in Table 2.

These parameters show a better correlation matrix (correlation matrix determinant=1.33) for using PCA algorithm. If PCA methods is applied then, it is shown that three principal components explain more than 75% (79,7%) of the variance and the matrix component should be Table 3.

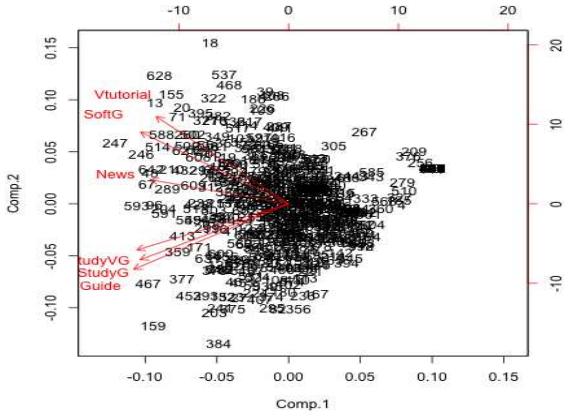

Figure 2: Data Set Draws in the PCA1 and PCA2 Spaces

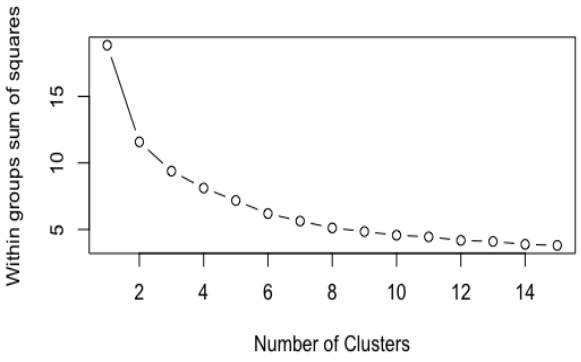

Figure 3: Elbow Algorithm for Determining the Number of Clusters (sum of squares)

The first component PCA1 it can be understood as the component that highlights the student that read all the previous information and follows the news. This kind of student it can be understand as a conscientious student. By other hand the second component describes the students are more interested in obtaining the more practical knowledge about the course: software guides and video tutorials, it may be considered as a practical student. Finally, the PCA3 describe a student that only is interested in what is happening in the course, probably is a sophomore or a student that do not understand the distance methodology. Since the highest variance is for the first two PCA components, (PCA1, PCA2) if it is drawn the data set in the space defined by those PCA components, it can be seen how the six parameters are related. Figure 2 represents the data set and the parameter vectors in the PCA1 and PCA2 space.

Taking into account these information, the next step was to classify the students to determine the possible states for a trait. For this purpose, a non-supervised learning algorithm has been chosen: K-means. To determine the number of possible group elbow method has been used. The result is in Figure 3. For this example, the number of the groups chosen was four.

One the number of groups is chosen the application of the k-means algorithm is straight and the result is shown in Figure 4.

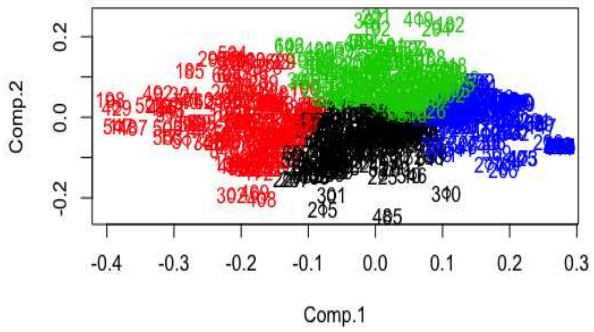

Figure 4: Elbow Algorithm for Determining the Number of Clusters (clusters)

As it was expected four group of students are built based on the PCA1 and PCA2 components. The four groups then can be described as:

- Group 1 (blue): PCA1 high. Students that read everything and prepare themselves to begin the course.
- Group 2 (red): PCA1 low. Students that neither read some much information nor are prepared to study.
- Group 3 (green): PC1 medium and PC2 High. Students that are more interested in quick and visual information and do not take care of another information maybe because they are sophomores.
- Group 4 (black): PC1medium and PC2 Low. Students that are neither interested in quick information maybe because they are sophomores.

In this way them it is possible to define the different behavior traits of a student and to fit them defining a number of states (groups) for each trait. Each student will be defined for a group of traits states and then will be possible to identified different behaviors.

4 CONCLUSIONS

In this work, the classification of students' learning behavior has been performed by proposing five different traits: general information of the course, learning resources, assessments, interactions tools, and implication of students. These five traits are composed of an amount of 71 indicators, as students' features. In order to reduce these parameters to study the most representative ones, a

PCA algorithm has been implemented, so reducing the dimensionality of our data set. An example of this reduction is given for one of the traits.

As future work, the PCA method will be used to all trait trying to define the students' behavior. By other hand, also it is possible to follow the same methodology using a new parameter: time. This allows us to analyze dynamically the different states of each trait for each student. The knowledge obtained will give insight about how the contents and the dynamic of the course influence in the student behavior. Also, the relation among traits will be studied, in order to check their correlation and define another more refined trait classification for the analysis of the students' behavior.

Finally, the authors' goal is to implement a system be able to determine automatically the students' behavior in order to improve the teaching-learning processes.

5 ACKNOWLEDGEMENT

Authors would like to acknowledge the support of the European research project ERC-2015-STG-679528 POSTDATA, the local project (2014I/PPRO/031) from UNED and Banco Santander and the Region of Madridor the support of E-Madrid Network of Excellence (S2013-ICE2715). The authors also acknowledge the support of SNOLA, officially recognized Thematic Network of Excellence (TIN2015-71669-REDT) by the Spanish Ministry of Economy and Competitiveness

REFERENCES

[1] 2017. Moodle logging 2. Web page at https://docs.moodle.org/dev/Logging_2. Date of last access: September 12, 2017. (2017).

[2] A. Daveedu-Raju, M. N. Sri, and G. L. Devi. 2016. Attribute subset selection by mixed weighting mean classification method. In *Intl. Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*. Chennai, India.

[3] C. Ding. 2005. Principal Component Analysis and Matrix Factorizations for Learning. Web page at http://ranger.uta.edu/~chqding/PCAtutorial/PCA-tutor1.pdf. Date of last access: September 12, 2017. (2005).

[4] S. L. S. López, R. P. D. Redondo, and A. F. Vila. 2016. Is interpersonal participation relevant to pass?. In *Fourth Intl. Conference Technological Ecosystems for Enhancing Multiculturality (TEEM)*. Salamanca, Spain.

[5] M. A. Mercado-Varela, A. G. Holgado, F. J. García-Peñalvo, and M. S. Ramírez Montoya. 2016. Analyzing navigation logs in MOOC: a case study. In *Fourth Intl. Conference Technological Ecosystems for Enhancing Multiculturality (TEEM)*. Salamanca, Spain.

[6] J. A. Ruipérez-Valiente, P. J. Muñoz-Merino, H. J. P. Díaz, J. S. Ruiz, and C. Delgado-Kloos. 2017. Evaluation of a learning analytics application for open EdX platform. *Computer Science and Information Systems* 14, 1 (2017), 51–73.

[7] M. Scheffel, H. Drachsler, S. Stoyanov, and M. Specht. 2014. Quality Indicators for Learning Analytics. *Educational Technology & Society* 17, 4 (2014), 117–132.

[8] Scikit-learn. 2017. Decomposing signals in components (matrix factorization problems). Web page at http://scikit-learn.org/stable/modules/decomposition.html. Date of last access: September 12, 2017. (2017).

[9] G. Siemens. 2013. Learning Analytics. *American Behavioral Scientist* 57, 10 (2013), 1380–1400.

[10] L. Tobarra, S. Ros, R. Hernández, A. Robles-Gómez, R. Pastor, A. C. Caminero, J. Cano, and J. Claramonte. 2017. Studying Students' Behavior in UNED-COMA MOOCs. In *Learning Analytics Summer Institute (LASI) Spain*. Madrid, Spain.