Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 155 (2019) 587–592

Procedia Computer Science

www.elsevier.com/locate/procedia

International Workshop on Applying Data Mining Techniques to E-Learning and Pedagogical Approaches (ADMEPA)
August 19-21, 2019, Halifax, Canada

Determinating Student Interactions in a Virtual Learning Environment Using Data Mining

Amelec Viloria[a]*, Jorge Rodríguez López[b], Karen Payares[c], Carlos Vargas-Mercado[d], Sonia Ethel Duran[e], Hugo Hernández-Palma[f], Mónica Arrozola David[g]

[a,c] *Universidad de la Costa, Street 58 # 55 - 66. Barranquilla, Colombia*
[b] *Universidad Simon Bolivar, Street. 58 #55-132, Barranquilla, Colombia*
[d] *Corporación Universitaria Latinoamericana, Street 58 #55 -24a, Barranquilla, Colombia*
[e] *Fundación Universitaria Unicolombo Internacional, Street 50 #31, Cartagena, Colombia*
[f] *Universidad del Atlántico, Street 30 # 8- 49 ,Puerto Colombia – Colombia*
[g] *Universidad Libre Seccional Barranquilla, Street 46 No. 48-170 , Barranquilla, Colombia*

Abstract

This article focuses on determining the students´ interactions in the Virtual English Course with Distance Education Model (DEM) at Mumbai University, in India. For this purpose, an analysis was carried out on the database of the students during the academic period 2015 - 2018 to select the necessary attributes that allowed to generate a data mining model. An analysis of the mining methods was subsequently carried out comparing each of them in order to select the one that helps the development of the project, choosing the Crisp-dm method since it contains multiple phases indicating each activity to be completed, thus becoming a practical guide. In addition, a comparative analysis was developed taking into account features of the data mining tools where RapidMiner was selected to perform the processes using some algorithms along with the student data which were divided into two sets for training and validation, resulting the decision tree as the best algorithm for the purpose as it correctly classified the instances with a minimum margin of error.

Keywords: data mining; classification technique; model; algorithm; Methodology.

* Amelec Viloria. Tel.: +57-3046238313
E-mail address: aviloria7@cuc.edu.co

10.1016/j.procs.2019.08.082

1. Introduction

Education is the base of any country's development, so today education systems around the world face the challenge of using information technologies which play an important role because they facilitate learning in environments allowing students the acquisition of immediate and broad knowledge without distance or time problems in their education [1], [2].

In this regard, the Mumbai University in India has information systems to provide distance education in careers or courses. These systems store large amounts of student information as is the case of the Virtual English course with a Distance Education Model, which was taken as the study object. However, having this information available could be a problem if there is not a method to process it, so data mining techniques are applied to extract useful and understandable knowledge, previously unknown, and discover patterns to generate a model through the analysis of information from course interactions, personal, institutional and student's socioeconomic data, to allow the determination of student interactions in the virtual course, so that it helps decision-making, and therefore provide a benefit to the institution [3], [4], [5].

Crisp-dm is one of the most used method for the generation of data mining projects, which allows to obtain a data analysis model along with the implementation of artificial intelligence algorithms, already integrated to the RapidMiner data pre-processing [6], [7].

2. Method

The data collected pertains to the student interactions in the Virtual English Course with the Distance Education Mode at the Mumbai University in India during the academic period 2015 – 2018. For this purpose, the information was obtained from the Moodle platform of the DEM, providing information about the interactions of the students in the following activities developed in the course: Interaction with shared files of course topics, tasks and evaluations for the approval of the course, reading or printing the course content and activities, send the activities to the teacher for correction, receiving qualifications, and online assessments and qualifications.

In this way, students are offered the opportunity to reinforce the learning provided through content and assessments to enable student-teacher and student-tools interactions.

The data is structured in XML files consisting of the interactions, personal, institutional, and socioeconomic data of the students, such as: number of accesses to the course, number of accesses to the tasks, number of accesses to resources, number of accesses to tests, description of modules, personal data of students, socioeconomic and institutional data of students [8], [9].

The generation of the test plan consists of testing the quality and validity of the results obtained by the model, therefore it is necessary to generate a test plan by which it is possible to test the validity of the generated model. So, the research selected the data belonging to students of DEM who were divided into two groups, one for training and the other one to use in the validation of the model.

The training dataset is 67% and the remaining dataset was used to perform validation in such a way that it gives a 100% result of the data used for modeling. The test plan is described with the different algorithms classified as follows [10], [11], [12], [13]:

- Decision Rules Algorithms: The algorithms used within this classification correspond to JRip, Ridor, Prism, K-NN, where 67% correspond to the training dataset and 33% were used for validation.
- Decision Tree Algorithms: The algorithms used within this classification correspond to CHAID, Decision Tree, ID3, J48, where 67% correspond to the training dataset (E) and 33% for validation (V). Parameters that were taken into account for evaluating the generated models are: correctly classified instances (accuracy), incorrectly classified instances (classification error), Kappa statistics which measures the match of the prediction with the actual class (Kappa), quadratic error, relative error, absolute error, presenting the obtained results in Table 1.

Table 1. Algorithm Results.

Algorithm	*Data*	*Quadratic Error*	*Relative Error (%)*	*Quadratic Error Middle*	*Quadratic Error Relative*
DECISION TREE	E	0.11	21.44	0.33	1.18
	V	0.07	13.33	0.25	4.29
JRip	E	0.05	10.18	0.23	0.80
	V	0.07	13.17	0.25	0.98
RIDOR	E	0.10	10.34	0.32	1.14
	V	0.11	11.34	0.31	1.23
K-NN	E	0.01	1.26	0.11	0.39
	V	0.16	15.85	0.39	1.51
PRISM	E	0.02	1.54	0.12	0.46
	V	0.26	26.06	0.50	2.37
Chaid	E	0.06	12.41	0.25	0.89
	V	0.17	27.35	0.41	5.46
ID3	E	0.01	1.97	0.09	0.39
	V	0.13	13.03	0.36	0.35
J48	E	0.08	16.07	0.29	1.01
	V	0.08	14.50	0.27	1.06

Table 1 shows the result of each algorithm obtained by using the RapidMiner tool along with student data from the Virtual English Course with the MED, where there is a minimum percentage of classification error in each of the algorithms. It can also be indicated that, with the training set of the data, most of the results obtained from the algorithms are favorable, i.e. they exceed 90% of the data were correctly classified, and the algorithms with better results are JRip 94.41%, K-NN 98.74%, Prism 98.46%, Chaid 91.06%, ID3 98.32, and J48 91.06%. Likewise, with the validation dataset, the algorithms that present the best results of correctly classified data were the Decision Tree 92.90%, JRip 92.63%, and J48 91.49%.

3. Results and Analysis

In order to determine student interactions, the best result of the algorithms was considered, which were analyzed in the evaluation of the model (Table 1) obtaining the Decision Tree algorithm as the best result, presenting a good classification with 92.9% and lower error margin in the model validation with 7.1%.

3.1 Rules Obtained Through Data Mining Algorithms

3.1.1 High level of interaction in the virtual course

When students have high interactions with exams and resources and are between 25 and 29 years old and single, then virtual course interactions are high.

If interactions with resources are average and between 25 and 29 years old and the gender is female, then interactions in the virtual course are high.

If exam interactions are average and they are older than 29 and don't work, then interactions in the virtual course are high.

If interactions with exams and resources are average and belong to another city and have no children, then interactions in the virtual course are high.

If test interactions are average and interactions with resources and tasks are high, and gender is female and do not work, belong to another city, and older than 29 years old, then interactions in the virtual course are high.

Interactions with exams and resources are high and have no service, have no children, and are between 25 and 29 years old, so interactions in the virtual course are high.

3.1.2 Average level of interaction in the virtual course

Exam interactions are average, belong to another city, have no children, and are married, so interactions in the virtual course are average.

If exam interactions are average, interactions with resources and tasks are high, have all services, do not work, gender is male, belongs to another city, and is older than 25 years, then interactions in the virtual course are average.

If interactions with resources are high, interactions with tasks are medium, gender is male, have all services, and is over 29 years old, then interactions in the virtual course are average.

The student does not work, the gender is male, belongs to another city, is single, and only has a service that is cell phone number, then interactions in the virtual course is medium.

The student does not work, the gender is female, has all the services, belongs to the city of Mumbai, is single, is under 25 years old, and interactions with resources is average, then interactions in the virtual course is average.

Exam interactions are low, interactions with tasks are low, interactions with resources are average, and married, so interactions in the virtual course are medium.

Interactions with exams and homework and resources are low, have services, have children, are over 29 years old, work, gender is male, belongs to another city, and is married, so interactions in the virtual course is average.

Exam interactions are average, interactions with tasks and resources are low, have all services, and are single, so interactions in the virtual course are average.

3.1.3 Low level of interaction in the virtual course

If interactions with exams and resources or tasks are low, and work, then interactions in the virtual course are low.

If interactions with exams and resources are low, gender is female, and has children, then interactions in the virtual course are low.

If interactions with exams and resources are low, have children, and are under 25 years old, then interactions in the virtual course are low.

If interactions with exams, tasks and resources are low, belongs to the city of Mumbai, does not work, has all services, gender is male, is over 29 years old, and has children, then interactions in the virtual course is low.

The student does not work, the gender is male, belongs to the city of Mumbai, has a service, is less than 25 years, and interactions with resources and tasks is low, then interactions in the virtual course is low.

The student works, the gender is male, belongs to the city of Mumbai, has all services, is single, is under 25 years old, interactions with resources and tasks is low, and has children, then interactions in the virtual course is low.

The student works, the gender is female, possesses all the services, belongs to another city, is married, and interactions with the exams is low, then interactions in the virtual course is low.

Interactions with exams and homework and resources are low, have services, have children, are older than 29 years old, work, gender is female, belongs to the city of Mumbai, and is widower, so interactions in the virtual course is low.

3.2 Factors to Determine Student Interactions

The factors that influenced the developing of the model are associated with each other [14], [15], considering personal, institutional, socioeconomic, and student interactions which are detailed below:

Course interactions: interactions/tasks (number of accesses to tasks), interactions/resource (number of times they access a resource), interactions/tests (number of test accesses).

Personal data of students: gender, marital status, age, services (phone, cell phone), city (students residing in Mumbai or in another city of the country).

Socioeconomic data of students: number/Children, work (whether the student works or not). Student institutional data: career (student's career).

Each attribute is presented below with its respective weights according to the results obtained by the Decision Tree algorithm to determine the highest influence on the model (see Table 2).

Table 2. Percentage of factors, attributes with their respective weights

Attribute	*Attribute percentage (%)*
Task interactions	12,196
Resource interactions	10,946
Exam interactions	13,299
Gender	4,562
Marital status	9,509
Age	8,671
Services	8,299
Career	9,137
Number of children	5,346
Work	8,126
City	9,908

4. Conclusions

Through the results, the following conclusions were obtained: Data mining is very important within the field of education as it helps extract information that is hidden in the data in such a way that it allows the analysis and generation of new knowledge to determine the level of student interactions. RapidMiner is a powerful data mining tool since it contains add-ons that allows to use different algorithms in this and other tool with operators that help in the development of processes of creating the applicable models for data analysis.

To determine the level of interaction in the English course, different classification algorithms were applied, presenting the best results in the Decision Tree since this algorithm obtained the least error margin during the classification of the data from the interactions in the course (tasks, tests, resources), and personal, institutional and socioeconomic data. Through this data mining model, it was possible to determine that student interactions in the English Virtual Course resulting the average level with a percentage of 69%, and the factors that influenced the model the most were the interactions in the student's exams, tasks, resources, marital status, and employment status.

References

[1] Viloria A., Lis-Gutiérrez JP., Gaitán-Angulo M., Godoy A.R.M., Moreno G.C., Kamatkar S.J. (2018) Methodology for the Design of a Student Pattern Recognition Tool to Facilitate the Teaching - Learning Process Through Knowledge Data Discovery (Big Data). In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham.

[2] Ballesteros Román, A.: Minería de Datos Educativa Aplicada a la Investigación de Patrones de Aprendizaje en Estudiante en Ciencias. Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada, Instituto Politécnico Nacional, México City (2012).

[3] Ben Salem, S., Naouali, S., Chtourou, Z., 2018. A fast and effective partitional clustering algorithm for large categorical datasets using a k-means based approach. Comput. Electr. Eng. 68, 463–483. https://doi.org/10.1016/j.compeleceng.2018.04.023.

[4] Chakraborty, S., Das, S., 2018. Simultaneous variable weighting and determining the number of clusters—A weighted Gaussian means algorithm. Stat. Probab. Lett. 137, 148– 156. https://doi.org/10.1016/j.spl.2018.01.015.

[5] Abdul Masud, M., Zhexue Huang, J., Wei, C., Wang, J., Khan, I., Zhong, M., 2018. Inice: A New Approach for Identifying the Number of Clusters and Initial Cluster Centres. Inf. Sci. (Ny). https://doi.org/10.1016/j.ins.2018.07.034
[6] Rahman, M.A., Islam, M.Z., Bossomaier, T., 2015. ModEx and Seed-Detective: Two novel techniques for high quality clustering by using good initial seeds in K-Means. J. King Saud Univ. - Comput. Inf. Sci. 27, 113–128. https://doi.org/10.1016/j.jksuci.2014.04.002
[7] Rahman, M.A., Islam, M.Z., 2014. A hybrid clustering technique combining a novel genetic algorithm with K-Means. Knowledge-Based Syst. 71, 345–365.
[8] Ramadas, M., Abraham, A., Kumar, S., 2016. FSDE-Forced Strategy Differential Evolution used for data clustering. J. King Saud Univ. - Comput. Inf. Sci. https://doi.org/10.1016/j.jksuci.2016.12.005
[9] Yaqian, Z., Chai, Q.H., Boon, G.W., 2017. Curvature-based method for determining the number of clusters. Inf. Sci. (Ny). https://doi.org/10.1016/j.ins.2017.05.024
[10] Tîrnăucă, C., Gómez-Pérez, D., Balcázar, J.L., Montaña, J.L., 2018. Global optimality in k-means clustering. Inf. Sci. (Ny). 439–440, 79–94. https://doi.org/10.1016/j.ins.2018.02.001
[11] Xiang, W., Zhu, N., Ma, S., Meng, X., An, M., 2015. A dynamic shuffled differential evolution algorithm for data clustering. Neurocomputing. https://doi.org/10.1016/j.neucom.2015.01.058
[12] Torres-Samuel M., Vásquez C.L., Viloria A., Varela N., Hernández-Fernandez L., Portillo-Medina R. (2018) Analysis of Patterns in the University World Rankings Webometrics, Shanghai, QS and SIR-SCimago: Case Latin America. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham.
[13] Vásquez, C., Torres, M., Viloria, A.: Public policies in science and technology in Latin American countries with universities in the top 100 of web ranking. J. Eng. Appl. Sci. 12(11), 2963–2965 (2017)
[14] Torres-Samuel M, Carmen Vásquez, Amelec Viloria, Tito Crissien Borrero, Noel Varela, Danelys Cabrera, Mercedes Gaitán-Angulo, Jenny-Paola Lis-Gutiérrez. (2018). Efficiency Analysis of the Visibility of Latin American Universities and Their Impact on the Ranking Web. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham.
[15] Piotrowski, A.P., 2017. Review of Differential Evolution population size. Swarm Evol. Comput. 32, 1–24. https://doi.org/10.1016/j.swevo.2016.05.003