2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)

Investigation of College Dropout with the Fuzzy C-Means Algorithm

Mariana Macedo*, Clodomir Santana Jr.†, Hugo Siqueira‡, Rodrigo L. Rodrigues§
Jorge Luis C. Ramos¶, Joao Carlos S. Silva¶, Alexandre Magno Andrade Maciel†, and Carmelo J. A. Bastos-Filho†
* University of Exeter, Exeter, UK
†University of Pernambuco, Recife-PE, Brazil
‡Federal Technological University of Parana - Ponta Grossa-PR, Brazil
§ Rural Federal University of Pernambuco, Recife-PE, Brazil
¶ Federal University of Vale do Sao Francisco, Petrolina-PE, Brazil
E-mails: mg615@exeter.ac.uk, cjsj@ecomp.poli.br, hugosiqueira@utfpr.edu.br, rlr@cin.ufpe.br,
jorge.cavalcanti@univasf.edu.br, jsedraz@gmail.com, {amam,carmelofilho}@ecomp.poli.br

***Abstract*—Up to 50% of the students drop out of school in Brazilian universities. Because of the heterogeneity of individuals, it is difficult to determine which are the main causes of this high percentage of students not finishing their degree. In this paper, we employed the Fuzzy C-Means algorithm on a dataset composed of real-world registers of the Biology Undergraduate course from Brazilian universities. We applied the transactional distance theory to select the set of variables which were utilized in the clustering process. The results indicate that the data is better divided into five groups. We observed that the Fuzzy C-Means generated groups based on how engaged the students are, and, in each group, there are two subgroups: students that drop out and do not drop out the course. The type of analysis presented in this work can generate inputs for the institutions to establish new policies to reduce the dropout rate.**

***Index Terms*—Education Assessment, School Dropout, Clustering, Distance Education, Gap Statistic**

I. Introduction

According to the Brazilian Association for Online Education (ABED), the school dropout rates reported in undergraduate online courses are higher than those in regular courses. The 2015 Census registered an evasion between 26% and 50%, with 40% of occurrences in institutions offering fully regulated courses [1]. Additionally, we highlight that this is not an issue restricted to Brazilian education institutions. Zawacki-Richter et al. reported graduation rates of 82% for students on full-time courses and around 39% for part-time students in universities from the United Kingdom [2].

Furthermore, the school dropout problem is related to several aspects such as the infrastructure, available tools, contents of the course, student motivation, among others. The 2017 Census on Distance Education published by the ABED, indicates that 59% of the credentialed institutions in Brazil do not know the reasons which are causing the elevated dropout rates in their courses [3].

The application of methodologies, such as the Knowledge Discovery in Databases, allows the analysis and extraction of information from the data related to courses dropout. Specifically, clustering techniques provide homogeneous groups according to the similarities presented on a dataset. As we usually do not have any a priori classification of users from online platforms, clustering is an unsupervised option [4] to unveil users peculiarities, and Fuzzy C-Means (FCM) [5] [6] is the simplest and fastest option that considers overlap data.

In this work, we applied the FCM algorithm into a dataset containing dropout information of students from an undergraduate course in Biology. The attributes used during the creation of the clusters were selected using the transactional distance theory, and we analyzed the quality of the clusters generated of students using the Gap Statistics metric. The primary goal of this study is to use the clustering process to generate groups of students whose characteristics might help to understand the reasons for dropout among the students and their profiles.

II. Background

The Fuzzy C-means (FCM) is a highly clustering algorithm, as it has the capability to find good partitions even when the data is overlapped; present simple mathematical treatment and fast convergence [7] [8] [9] [10].

Since the FCM algorithm considers overlapping data, it does not define a hard partition of K clusters. Instead, the method creates a membership matrix $\mathbf{U}$, having as the inputs the membership elements μ_{ij} of the patterns i for each cluster j. This matrix determined by $\mathbf{U} = \{\mu_{ij}\}_{i,j}^{N,K}$, being $\mu_{ij} \in [0,1]$, $\sum_{j=1}^{K} \mu_{ij} = 1$, $\forall i$, and $0 < \sum_{i=1}^{N} \mu_{ij} < N$. The goal is to optimize the objective function given by Equation 1, in which $||.||$ represents the Euclidean norm. Then, the iterative process occurs in two steps: i) it updates μ_{ij} for each i and j using Equation 2, in which C is the maximum number of clusters. Then, it updates the centers' positions according to Equation 3, where m is the fuzziness coefficient provided by the user.

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{K} \mu_{ij}^m ||\mathbf{x}_i - \mathbf{c}_j||^2, \tag{1}$$

$$\mu_{ij}^m = \frac{1}{\sum_{k=1}^{C} \frac{||\mathbf{x}_i - \mathbf{c}_j||}{||\mathbf{x}_i - \mathbf{c}_k||}^{\frac{2}{m-1}}}, \tag{2}$$

2161-377X/19/$31.00 ©2019 IEEE
DOI 10.1109/ICALT.2019.00055

$$\mathbf{c}_j = \frac{\sum_{i=1}^{N} \mu_{ij}^{m} . \mathbf{x}_i}{\sum_{i=1}^{N} \mu_{ij}^{m}}. \quad (3)$$

Tibishirani et al. introduced the Gap Statistics (Gap) metric as an alternative to determining the correct number of clusters in a database [11]. Some studies showed the efficiency of the method in comparison with some popular metrics [11]. The GAP compares a randomly generated dataset with the one in the study; the statistical difference from the expected shows the quality of the clustering [11]. Moreover, GAP measures the distance of all two combinations of samples that belong to each cluster r. In summary, GAP is the subtraction of the expectation (E_n) of the logarithm of W, to both sets.

III. Educational platform profiles

To perform this analysis, we address databases of LMS (Learning Management System) Moodle of the Nucleus of Online Education (NEAD) - University of Pernambuco (UPE). The NEAD has about 10 years of activities in online education, offering undergraduate, specialization and extension courses for several cities within the state of Pernambuco, Brazil.

We chose the database of the Biological Sciences undergraduate course, in the online education. We collected the data of the eight semesters (entire course) regarding the students' performance and activities. Thus, we used the data of all the subjects attended by the regular students of the course. This database contains a total of 1,150 instances.

The university provided the samples in SQL format. Over the years, it adopted both PostgreSQL and MySQL as DBMS, as well as different versions of Moodle. The consistency checks, on the obtained data, were made through an administrator access in the LMS Moodle. For example, we compared the records of the number of specific interactions that the reports of the own environment made available, with the data collected by the SQL scripts. We performed other checks crossing different queries and comparing their returns with the already collected values.

For the execution of the procedures of this step, we used the MySQL Workbench and pgAdmin III applications to extract the data through SQL scripts; we deployed the statistical package R with its RStudio IDE for performing a preliminary analysis and data processing, and we also used Microsoft Excel for treating and standardizing the data.

For each student, we selected fourteen (14) variables related to the constructs of the Transactional Distance Theory: dialogue, structure, and autonomy [12]. These constructs were developed in [1], being representative of a predictive model of drop out in online education. The list with a short description of this variables is available in http://bit.ly/2OAKyJF .

IV. Case Study and Analysis

In this section, we named the students that dropped out the Biology course as *temporary* students and the others that did not drop out based on the database period are named *permanent* students. From our experiments, we observed that the structural characteristics did not show any difference through the biology students. We believe that further structural characteristics should be extracted by the platform in other to enhance this constructor in the educational platform.

Figure 1 shows that the number of profiles we can observe for the Biology course is equal to 5. Moreover, Table I shows the percentage of temporary and permanent students, and Table II displays the average value of each cluster's feature.

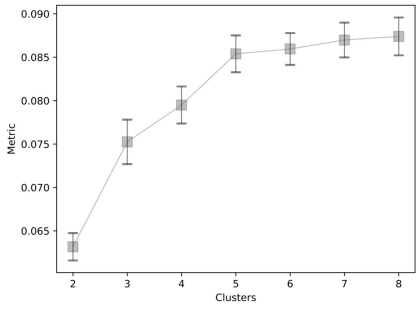

Fig. 1. Results of the GAP metrics as a function of the number of clusters.

TABLE II
Summary of clusters

Cluster	Instances	Temporary	%	Permanent	%
#0	74	14	18.9%	60	81.1%
#1	266	118	44.4%	148	55.6%
#2	173	42	24.3%	131	75.7%
#3	98	13	13.3%	85	86.7%
#4	171	34	19.9%	137	80.1%
Total	782	221	28.3%	561	71.7%

In cluster label **#0**, all the best students are identified which explains the number of temporary students being low. The permanent students from this group use the framework in the manner it is supposed to be. Temporary and permanent students are easily divided into groups. These students seem to be dedicated, but for some obscure reason they got "frustrated". Options of harder materials or exercises for this group may be important to maintain the dedicated students inspired to continue the course.

In cluster label **#1**, we can observe the highest number of temporary students, and autonomy has being differed from more than 50% of each feature. This group should receive more notifications reminding them to accomplish their tasks because they do not use sufficiently the platform. Another solution is the gamification of activities for this group, so they become more excited to finish their tasks. Further analysis in this group may give other insights into how ineffective the platform is for them.

Cluster label **#2** shows to be the hardest group to understand because both temporary and permanent students use the platform similarly. However, the dialogue has less similar characteristics. The necessity of more dialogue features will establish more differences in this group.

The highest percentage of permanent students is displayed in cluster label **#3**. They are divided by the daily number of

TABLE I
AVERAGE VALUES OF THE RESULTS.

CLUSTER 0	**var01**	**var02**	**var03**	**var04**	**var10**	**var12**	**var13**	**var16**	**var17**	**var20**	**var21**
Temporary	8.21	58.50	73.00	75.50	9.24	60.64	5.14	35.50	130.50	19.00	15.50
Permanent	12.25	118.43	135.42	60.97	15.49	99.67	7.10	69.12	139.85	23.53	25.25
CLUSTER 1	**var01**	**var02**	**var03**	**var04**	**var10**	**var12**	**var13**	**var16**	**var17**	**var20**	**var21**
Temporary	0.61	3.93	4.52	7.08	1.16	2.54	0.46	1.41	22.11	0.11	0.14
Permanent	1.75	10.59	15.82	18.48	3.14	10.18	2.06	6.38	36.49	0.13	0.03
CLUSTER 2	**var01**	**var02**	**var03**	**var04**	**var10**	**var12**	**var13**	**var16**	**var17**	**var20**	**var21**
Temporary	2.77	15.88	20.71	33.38	5.42	25.71	3.57	2.69	56.88	0.24	0.12
Permanent	2.95	18.74	24.63	31.77	5.73	27.28	4.00	8.14	50.87	0.74	0.35
CLUSTER 3	**var01**	**var02**	**var03**	**var04**	**var10**	**var12**	**var13**	**var16**	**var17**	**var20**	**var21**
Temporary	4.18	10.54	33.08	62.31	10.91	60.62	6.31	26.92	105.69	10.38	8.92
Permanent	6.29	44.84	55.91	61.03	9.99	59.29	5.63	25.43	90.53	4.16	4.90
CLUSTER 4	**var01**	**var02**	**var03**	**var04**	**var10**	**var12**	**var13**	**var16**	**var17**	**var20**	**var21**
Temporary	3.51	17.09	24.47	47.06	8.72	40.09	4.26	9.71	82.06	5.62	3.32
Permanent	4.72	32.46	40.81	46.86	6.92	37.04	4.42	16.69	62.18	0.97	1.36

access and the number of communications. This group presents dedicated students. The temporary students probably may have the perception that the course is quite easy or probably they would like to have more advanced materials, since, indeed, the problem was not the lack of usage on the platform.

Cluster label **#4** has the people that work the whole day and only access the course by night. They are dedicated to trying to communicate and access the materials, but they do not deliver much. This group may need more time to perform better. The platform should offer longer courses for them.

Finally, we argue that using clustering to understand the students' profile better is relevant because it helps the platform to enhance the functionalities for each type of group. We found that more features should be analyzed regarding each aspect: Autonomy, Dialogue, and Structure. Moreover, the results showed different reasons for student dropout, which means that our personality and the amount of free time are essential. The establishment of specific functionalities for each group on the platform may decrease the number of temporary students. For example, the platform, using a dynamic decision tree, can continuously classify the users in one of the five groups and offer additional notifications, inspirations, and materials when necessary.

V. CONCLUSIONS

The application of a clustering method to investigate the reasons for school dropout brings insights to enhance online educational platforms. The case study evolves a database formed by students of the Biology bachelor course at the University of Pernambuco. For each student, we used 14 attributes to group the considered students. The computational tool addressed was the Fuzzy C-means algorithm, and the GAP statistics defined the correct number of clusters. The results showed 5 clusters with different characteristics and requirements, and further functionalities can be implemented to improve the online education platform. As future works, we believe that new features from each constructor should be created, a comparison to the big five personality traits should be investigated about the encountered groups, and a survey of the impact of the implementation of new functionalities should be performed.

ACKNOWLEDGMENT

This study was financed in part by the Coordenao de Aperfeioamento de Pessoal de Nvel Superior - Brasil (CAPES) - Finance Code 001, FACEPE, under grant IBPG-0964-3.04/16 and CNPq, process number 405580/2018-5.

REFERENCES

[1] J. L. C. Ramos, A. S. Gomes, R. Rodrigues, J. Silva, F. d. F. de Souza, E. de Gouveia Zambom, and L. Prado, "Um modelo preditivo da evasão dos alunos na ead a partir dos construtos da teoria da distância transacional," in *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, vol. 28, no. 1, 2017, p. 1227.

[2] O. Zawacki-Richter and T. Anderson, "Educação a distância online: construindo uma agenda de pesquisa," *São Paulo: Artesanato Educacional*, 2015.

[3] A. B. de Educação a Distância (org.), "Censo ead.br 2017 - relatrio analtico da aprendizagem a distncia no brasil," *São Paulo*, 2018. [Online]. Available: http://cbic2017.org/papers/cbic-paper-93.pdf

[4] S. Alam, G. Dobbie, Y. S. Koh, P. Riddle, and S. U. Rehman, "Research on particle swarm optimization based clustering: A systematic review of literature and techniques," *Swarm and Evolutionary Computation*, vol. 17, pp. 1–13, 2014.

[5] B. J.C., *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, New York, 1981.

[6] A. A. A. Esmin and R. A. Coelho, "Consensus clustering based on particle swarm optimization algorithm," in *Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE Computer Society, 2013, pp. 2280–2285.

[7] Y. Yong, Z. Chongxun, and L. Pan, "A novel fuzzy c-means clustering algorithm for image thresholding," *Measurement Science Review*, vol. 4, no. 1, pp. 11–19, 2004.

[8] S. Chattopadhyay, D. K. Pratihar, and S. C. De Sarkar, "A comparative study of fuzzy c-means algorithm and entropy-based fuzzy clustering algorithms," *Computing and Informatics*, vol. 30, no. 4, pp. 701–720, 2011.

[9] A. Stetco, X. jun Zeng, and J. Keane, "Fuzzy c-means++: Fuzzy c-means with effective seeding initialization," *Expert Systems with Applications*, vol. 42, no. 21, pp. 7541–7548, 11 2015.

[10] S. Ghosh and S. K. Dubey, "Comparative analysis of k-means and fuzzy c-means algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 4, 2013.

[11] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J. R. Statist. Soc. B*, vol. 63, no. 2, 2001.

[12] M. G. Moore, "The theory of transactional distance," in *Handbook of distance education*. Routledge, 2013, pp. 84–103.