An investigation of students behavior in discussion forums using Educational Data Mining

Crystiano J. R. Machado, Bruno R. B. Lima and Alexandre M. A. Maciel
University of Pernambuco - UPE
Recife, Brazil
{cjrm, brbl, amam}@ecomp.poli.br

Rodrigo L. Rodrigues
University Federal Rural of Pernambuco
Recife, Brazil
rodrigo.linsrodrigues@ufrpe.br

***Abstract*— Discussion forums are an important feature in the Distance Education courses, supporting learning and facilitating interaction between students and teachers. This work aims to investigate behavioral aspects of students in virtual learning environments using Educational Data Mining. It is proposed to use the *K*-Means clustering technique as a way to identify students with common patterns of behavior based on their interactions in the forums. This work achieves good results from the application of clustering technique for this particular issue.**

Discussion Forums; Distance Education; Behavior Aspects; Educational Data Mining; Clustering.

I. Introduction

Demand for Distance Learning (DL) has shown growth rates increasing. The results of educational census conducted in Brazil between the years 2013 and 2014 show that the 309 institutions surveyed, there are a total of 4,044,315 students enrolled in distance learning courses. This number is expected to become even more significant in 2015, as 82% of these institutions enrollment will be even greater. These results guarantee the DL one leadership position among the types of education in Brazil [1].

Part of technological solutions used in distance education are contained in so-called Virtual Learning Environments (VLEs). According to Santana [2], the growth of distance education has motivated the academic community and boosted the development of scientific research. These studies are favored because of VLEs possess the ability to accumulate in your databases a lot of information that is valuable for the analysis of student behavior [3]. Majority of this information refers to the communication generated by the students through the existing tools in the DL environments such as discussion forums, chats and digital media postings.

According to Azevedo [5], the involvement of students in discussion forums is an important activity for building the knowledge. By analyzing the interaction of students in the forums, the teacher can diagnose relevant information about students. On the other hand, due to the large volume of data, the analysis of this information requires the help of specific tools.

In this context, a strategy widely used today is the Educational Data Mining (EDM). Bittencourt and Costa [4] define EDM as an emerging area of research that seeks to develop, adapt and apply methods for the discovery of knowledge, in order to identify data sets from large information bases in educational systems.

The objective of this study is to present an investigation developed with students, teachers and tutors of undergraduate courses in distance education and intends to examine behavioral patterns of students concerning interactions in the forums of the analyzed disciplines as a way to identify groups with common characteristics. Therefore, this paper is organized as follows: section II presents the Theoretical Foundation; in section III is presented in detail the process of the Experiments; Section IV Results and Discussions are presented; in Section V are held the Conclusions about this work.

II. Theoretical foundation

A. Educational Data Mining

According to Webber, Zap and Lima [7], the Educational Data Mining (EDM) is an academic field of research that seeks to obtain, from the data stored in educational environments, new and useful information in order to develop and strengthen the cognitive theories of teaching and learning. More comprehensively, the "knowledge discovery can help teachers conduct their classes better, identifying difficulties, understanding better the learning process of students and improving teaching methods" [8].

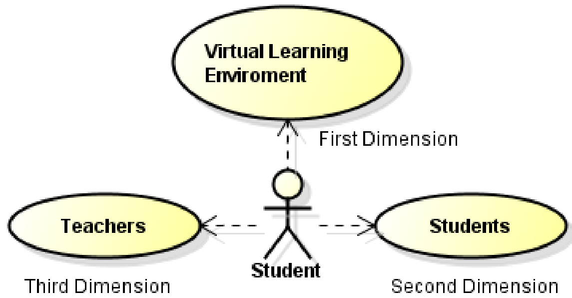
For Manhães, Cruz, Costa, Zavaleta and Zimbrã [9], the EDM is related to the application of traditional techniques Data Mining in the several data fields of application in educational settings and that mostly arise from the Distance Learning environments. However, due to the existence of the information on different hierarchy levels, there are algorithms and tools used in Data Mining area can’t be applied in the analysis of educational data without modification [10].

In parallel to the growth of research that is carried out in the area of EDM, so does the amount of data that is stored from the DL. For Romero, Ventura and García [8], this is due to inclusion of new features that are inserted in the Distance Learning platform. This enables interactions as: access to forums, questions and answers in areas that are designed to questionnaires and also due to direct communication between participants.

DOI reference number: 10.18293/SEKE2016-143

Gottardo, Kaestner and Noronha [6] propose that the representation of actions taken by a student in a Virtual Learning Environment (VLE) can be divided into three dimensions, seeking to contemplate the several aspects of use and interaction, as shown in Figure I.

FIGURE I. Representation of a Student Interaction in VLE.



- **First Dimension**: General profile of use of the VLE: this dimension aims to identify data representing characteristics of planning, organization and student time management for the course.
- **Second Dimension**: Student-Student interaction: the purpose of this dimension is to analyze characteristics that come to see if students interact with each other using the available tools of the VLE.
- **Third Dimension**: Student-Teacher interaction: this dimension aims to investigate how teachers or tutors interact with students in the context of VLE.

B. *Cluster analysis*

According to Baker [10], Educational Data Mining can be divided into five primary categories, one is called Clustering and aims to split the data into groups that share similar characteristics. This work focuses on the cluster analisys technique through *K*-Means algorithm.

According to Steinley [11], *K*-Means is one of the most used existing clustering methods in the literature, being defined as a partitioning method based on the definition of core elements of clusters, called centroids. Complementing this idea, [12] mentions that this partitioning occurs because the data set is divided into *K* subsets, where the value *K* is predetermined.

Among the main advantages of using the *K*-Means clustering technique as are its simplicity and speed. These characteristics favor its application in the process of mining of large databases. The disadvantage of the algorithm, one can consider the dependence of the parameter *K* represents the number of clusters on which data should be grouped together. In addition, the startup of the cluster centroids influences quality [12]. In order to minimize these negatives [13] suggests that the algorithm is executed with different values of *K* and boots.

According to Puma-Villanueva and Zuben [14], the main challenge in the realization of clusters is the analysis and definition of a great number of groups. Therefore, in this work, it adopted the technique Silhouette Index (SI) as a method of evaluation of clusters during the review process of the groups. The SI aims to evaluate and assign a grade to the quality of the result of the group, taking into account the distances inter and intra groups. Well separated and compact groups have better quality.

In implementing the SI on the formed group, each group observation presents a value that varies in the interval [-1, 1]. The closer the value of a note is to 1, the smaller the distance (or dissimilarity) between the analyzed observation and any comments allocated to other groups. Thus, it can be considered that the individual was properly allocated in the current cluster. Moreover, SI values close to -1 indicates that the individual is allocated to an unsuitable group. Values near 0 indicate that the individual does not belong to a cluster or another [15].

C. *Related works to this research*

Currently, VLEs have assumed a prominent position in the academic and corporate education, and because of the resources offered by these environments, has been made possible synchronous and asynchronous communication between teachers, tutors and students participating in the course of Distance Learning [16]. As a result, several scientific works related to the EDM have been conducted.

Some of these works aimed to estimate the academic performance of students, as in [17], where an analysis is made using mining techniques student data in a VLE, reaching levels between 73% and 80% success in the performance estimates of the students. Moreover, [18] used a linear regression model in order to estimate the performance of students based on their virtual platform of learning interactions, taking into account lifestyle factors.

[19] Conducted experiments using decision trees algorithms in order to identify early graduate students the distance with risk avoidance, allowing the discovery of the disciplines that influence more in the avoidance of an undergraduate degree. In [20] presents an analysis of improvement of pedagogical actions, applying techniques of MDE in order to support behavior detection processes linked to truancy.

From the point of view of the interactions in distance education forums, [21] use statistical techniques and EDM in order to investigate the pattern of behavior of students and teachers of distance learning courses, using association rules and classification algorithms. [22] Use the semi-supervised algorithm *SVM-KNN* with the proposal to create new possibilities for the management and automatic monitoring of the forums on distance education courses. In turn, [23] use clustering techniques in forums publications to generate indicators for Educational Management.

III. EXPERIMENTS

In this work, Moodle data were collected, the Learning Management System (LMS) used more between higher education instructions offering distance education courses in Brazil [1]. These data were obtained considering some attributes related to forms of student interaction in the discussion forums. The completion of the experiments was based on data from 46 students of Educational Organization and Public Policy course, of the course Physical Education. The result points relevant factors that should be considered in order to assist teachers and tutors in the specific monitoring needs of each student.

May be considered that such interactions arise from the exchange of information among students who share the same VLE to carry out a course as well as the exchange of information between students and their teachers or tutors. Moreover, it is also regarded as a type of interaction the actions that a student performs in the system, involving only the student and the VLE.

The selected attributes for this work were based on the study realized by Gottardo, Kaestner and Noronha [17], which were associated with various attributes of Moodle to each of the three students interaction dimensions in a VLE [6]. Table I shows the attributes discussed in this work during the process of Data Mining.

TABLE I. LIST OF ATRIBUTES (BASED ON [17]).

Dimension	Attribute
General Profile of Use of the VLE	1. Total number of posts held in forums.
	2. Number of posts of other participants read on forums.
	3. Total number of revisions in previous posts held in forums.
Student-Student Interaction	4. Number of answers posted in forums referring to postings from other students.
	5. Number of postings of other students doing student posting reference.
Bidirectional Interaction Student-Teacher	6. Number of student posts that had responses made by teachers or tutors of the course.
	7. Number of teacher posts or guardians who had responses made by the student.

A. Preprocessing

Starting the preprocessing of data, a two-dimensional table is built, containing the list of students associated with their respective values for each attribute. Then, the table was submitted to the preprocessing and data transformation, in order to improve the quality of the mining process. Such procedure involved the following steps:

- Analysis of raw data - In this stage, for each attribute, the statistical analysis of raw data collected from the database through SQL queries was held. This analysis was performed through the values processed by the Weka tool, so that they could be aware of the data set characteristics.
- Cleaning the data - this step was done filling the columns that did not have defined values for attributes. Instances in the table that did not have a value associated with a particular attribute, had zero value assigned by default.
- Identification of outlayers - In this stage, for each variable, the identification of instances in the table of each class was held, seeking to inconsistent values when compared with the other set of data. For this, each of the entries in the tables were compared with the standard deviation and the average value of the respective attribute.
- Removal of outlayers - After the identification of inconsistent entries in the tables for each attribute considered in this research, adjustments were made in SQL queries.
- Standardization - this step was carried out adjusting the scale of values for each attribute by linear normalization. The values used in this study corresponded to the interval [0,1].
- Analysis of the transformed data - Finally, the new set of data representing the attributes individually in each class was again assessed by Weka. The use of the tool at this stage of the transformation process enabled the interpretation consistent with values.

B. Organization of scenarios

The scenarios examined in this study are based on the variation of the number of groups depending on the size of the interactions of the students. During the execution of the experiments, we observed that the implementation of K-Means for formations from eight groups, generated clusters with only one student. Therefore, we considered scenarios with the number of groups ranging from two (minimum) and seven (maximum). As this technique the "K" represents the number of groups, we have K = {2, 3, 4, 5, 6, 7}.

For each value of K, scenarios were considered containing the attributes of each of the three dimensions. This approach allowed for the analysis of each dimension individually and therefore identify groups with specific behavior patterns. Therefore, considering the variation of attributes and also the number of clusters, They analyzed a total of 24 scenarios. Such scenarios can be divided into four sets.

The first group of analyzed scenarios aims to investigate behavioral aspects of the students related to the VLEs usage profile. The second group of scenarios discussed in this paper aims to investigate behavioral aspects of the students related to interactions between students in the forums. The third group of scenarios analyzed aims to investigate behavioral aspects of the students related to the two-way interactions between students and teachers or tutors in the forums. The fourth group of simulations aims to investigate students' behavioral aspects in the discussion forums, based on three dimensions simultaneously.

C. Organization of scenarios

The results with the formation of the groups were subsequently subjected to analysis of Silhouette Index (SI). The SI was implemented to receive as input the distribution of the groups formed by Weka and the data with the attributes of each class. As a result, SI displays in percent the quality of clustering formed by the *K*-Means.

Was selected the scenario that had better quality percentage to determine the best number of groups (value of *K*), i.e., the best formation of groups by analyzing all dimensions simultaneously. The best scenario was chosen for the analysis stage of groups and analysis of the behavioral aspects of students.

After identifying the best distribution of students according to the number of groups, it was possible to individually analyze each of the dimensions and, according to the attributes of each, draw a profile of groups based on their interactions.

IV. RESULTS AND DISCUSSIONS

The results obtained after the experiments are presented and discussed below. As a result of this research, analysis of the behavioral aspects of the students was based on the characteristics of the formed clusters.

For the 46 students of the discipline analyzed it was found that the best result in the formation of groups process, considering the set of attributes of the 1st, 2nd and 3rd dimension simultaneously, occurred with $K = 7$. In this scenario, the quality of clustering was estimated at 28,5%, as shown in Table II. These groups have respectively 1, 9, 3, 11, 2, 10 and 10 students.

TABLE II. CLUSTERING PROCESS QUALITY FOR EACH DIMENSION OF INTERACTION.

	Dimension			
	1st	**2nd**	**3rd**	**1st, 2nd e 3rd**
$K = 2$	36,34%	44,73%	41,95%	26,24%
$K = 3$	34,36%	45,91%	43,72%	21,63%
$K = 4$	27,44%	43,8%	41,4%	21,49%
$K = 5$	36,18%	39,63%	40,17%	25,93%
$K = 6$	30,49%	24,14%	37,95%	27,17%
$K = 7$	31,94%	18,76%	40,13%	28,5%

According to the statistical values related to the attributes for student's clusters were formed, it is possible to identify some important features for each group.

In Group 1, students are found with reasonable average of posts, with an average of 45,67, and satisfactory values regarding the amount of revisions posts, with an average of 15,17, representing students with second highest average in this attribute. Students in this group have a satisfactory profile of VLE usage and with good levels of interaction with each other. However, there is also the low amount of posts that are read by the students of this group, represented by average of 0,33 read messages. I notice also that in this group there is greater involvement by the students than teachers or tutors.

In Group 2 are the students with the highest average postings made in the forums and highest average revisions posts, with average respectively equal to 80,88 and 21,38. However, this group of students have values unsatisfactory as regards the number of read messages. Students of this group, as well as in Group 4, did not stand out in any of the characteristics related to the interactions student-student and Student-Teacher.

In Group 3, besides the unsatisfactory average of read messages, there is poor performance on the number of posts held and the lower middle revisions on posts, and the group with the worst results in both aspects analyzed. This group represents students with less interaction with distance learning platform and unsatisfactory levels of interaction with other students, teachers and tutors.

In Group 4 students with satisfactory average posts are found (average equal to 39,83), but with low average interaction through read or revised posts. In this group, none of the interactions of the students stands out compared to the other groups. In addition, the interaction of the students of this group in teachers or tutors posts is the second worst result, compared with the other groups.

Group 5 is characterized by representing students with more interaction with distance education platform. Furthermore, the Group 5 is students with good interaction with other students, being the group with the highest average in both attributes, represented by average equal to 83,75 and 63,50 to the attributes related to the Student-Student Interaction. You can also see the good level of interaction of these students with teachers or tutors of the course.

The class of Groups 6 and 7 have a satisfactory average as regards the number of posts, with an average respectively equal to 68,89 and 70, and resemble the amount of unread posts, with values unsatisfactory for this feature. However, such groups differ in the average revisions in posts made by students, where students of the Group 6 have satisfactory values for this attribute (average equal to 11,67), unlike the students in the Group 7 (average equal to 4,33). The Group 6 has the distinction of the good level of interaction with teachers and tutors of the course, with an average of 20 posts that had responses made by teachers or tutors of the course, and 53 posts who had responses made by the student.

The similarity of the behavioral pattern was found of students forming groups 3, 4, 5 and 6 with the results presented by Silva *et. al*. [24], where they investigated the actions of students in forums, chats and downloads using the Ward method. Our work is distinguished by applying the K-Means technique for training of students clusters in discussion forums and then validating the quality of the clusters through technical Silhouette Index. In addition, our study analyzes the interactions of students in three different dimensions, based on Gottardo, Kaestner and Noronha [17].

With the approach used in this study made it possible to obtain significant results in the identification of homogeneous groups of students (i.e., students of the same group have a relevant level of similarity on the aspects analyzed), as well as in-depth analysis of its features. The analysis of the formed clusters are useful for course coordinators and mentoring, as well as teachers and the own students, to the extent that point to important indicators for search of valuable experiences related to students' interactions in order to replicate them in other classes.

V. CONCLUSIONS

Discussion forums are an important feature in VLEs, for providing a support to the learning process that facilitates interaction by students, teachers and tutors in distance education environments. We can affirm that, through the forums, one can notice the individual participation of students, discussing the topics of the courses and contributing to the construction of collective knowledge.

Thus, this study aimed to conduct experiments using the K-Means clustering technique on data extracted from the Moodle forums on 46 students of Educational Organization and Public Policy course, of the course Physical Education in a higher education institution. The results were analyzed by the technique Silhouette Index, in order to identify the best training groups in their respective disciplines. Finally, the interpretation of the results to identify the behavioral aspects of the students was held, based on three dimensions of interaction: General Purpose Profile VLE, Student-Student Interaction and Student-Teacher Interaction Bidirectional.

Analyzing the students of Educational Organization and Public Educational Policy discipline, can form seven groups with very particular characteristics. Group 3 present unsatisfactory results in three dimensions, with students who have a very limited interaction profile. Some groups stand out in specific dimensions, the group 5 for the General Purpose Profile VLE interactions and the interactions student-students, and the group 6 over the Bidirectional Interactions Student-Teacher. The other groups of this course stand out in specific ways in each of the three dimensions.

The identification of groups of students from the interaction profile in the discussion forums are useful in order to provide teachers with an educational support that enables a deeper understanding of behavioral aspects of their students. The result of the identification of the characteristics of groups of students can help teachers plan their class better, identify the possible difficulties of students, better understand the learning process of students and improve teaching methods.

To obtain better results for future work, analyzes will be made that in addition to quantitative characteristics, incorporates qualitative aspects to obtain results that can infer possible learning problems of students, from the content of the interactions in the discussion forums. Another relevant aspect for future works concerns a proposal that incorporates assessment of external quality of the clustering, in order to identify an optimal clustering model.

References

[1] ABED. Associação Brasileira de Educação A Distância. Censo EaD.br: relatório analítico da aprendizagem a distância no Brasil 2013. 1 ed. Curitiba: Ibpex, 2014.

[2] Santana, L. C. "Integração de um mecanismo de mineração de dados educacionais ao moodle". 2015. 85p. Dissertação (Mestrado em Engenharia de Computação) - Escola Politécnica de Pernambuco - POLI, Universidade de Pernambuco - UPE, Recife. 2015.

[3] Mostow, J.; Beck, J.; Cuneo, A.; Gouvea, E.; Heiner, C. A Generic Tool to Browse Tutor-Student Interactions: Time Will Tell!. In AIED (pp. 884-886). Maio, 2005.

[4] Bittencourt, I. I.; Costa, E. B. (2011) "Modelos e Ferramentas para a Construção de Sistemas Educacionais Adaptativos e Semânticos". Revista Brasileira de Informática na Educação, v. 19, p. 85-98.

[5] Azevedo, B. F. T. "Análise das mensagens de fóruns de discussão através de um software para mineração de textos". Anais do XXII Simpósio Brasileiro de Informática na Educação, Aracaju. 2011b.

[6] Gottardo, E.; Kaestner, C.; Noronha, R. V. "Avaliação de Desempenho de Estudantes em Cursos de Educação a Distância Utilizando Mineração de Dados". Anais do XXXII Congresso da Sociedade Brasileira de Computação. 2012.

[7] Webber, C. G.; Zat, D.; Lima, Mª. de F. W. do P. Utilização de Algoritmos de Agrupamento na Mineração de Dados Educacionais. Revista RENOTE - Tecnologias na Educação. V 11. Nº. 1. Julho, 2013. ISSN 1679-1916.

[8] Romero, C.; Ventura, S.; García, E. Data mining in course management systems: Moodle case study and tutorial. Computers and Education, 51 (1), pp. 368–384. 2008.

[9] Manhães, L. M. B.; Cruz, S. M. S.; Costa, R. J. M.; Zavaleta, J.; Zimbrã, G. Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. Instituto de Ciências Exatas – Universidade Federal Rural do Rio de Janeiro (UFRRJ). 2011.

[10] Baker, R. S. J. D. Data Mining for Education. McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition). Oxford, UK: Elsevier. 2010.

[11] Steinley, D. Standardizing variables in K-meand clustering. In D. Banks, L. House, F. R. McMorris, P. Arabie, & W. Gaul (EDs.), Classification, clustering and data mining applications (p. 53-60). New York: Springer. 2004.

[12] Malta, C. S. "Estudos de Séries Temporais de vento Utilizando Análises Estatísticas e Agrupamento de Dados". Tese de mestrado. Rio de janeiro, RJ – Brasil, Fevereiro de 2009.

[13] Souza, T. A. "Agrupamento de Séries temporais de Vento para Avaliação da Disponibilidade de geração de Usinas Eólicas". Universidade Federal do Rio de Janeiro. UFRJ, COPPE. Dissertação de Mestrado. 2008.

[14] Puma-Villanueva, W. J.; Zuben, F. J. V. "Índices de validação de agrupamentos". Universidade Estadual de Campinas. Unicamp. Faculdade de Engenharia Elétrica e de Computação. FEEC. 2008.

[15] Anzanello, M. J.; Fogliatto, F. S. Selecting the best clustering variables for grouping mass-customized products involving workers' learning. International Journal of Production Economics 130 (2), 268-276. 2011.

[16] Maciel, A. M. A.; Rodrigues, R. L.; Carvalho, E. C. B. "Desenvolvimento de um Assistente Virtual Integrado ao Moodle para Suporte a Aprendizagem Online". In: III Congresso Brasileiro de Informática na Educação, 2014, Brasil. Anais do XXV Simpósio Brasileiro de Informática na Educação, 2014. p. 382-391. ISSN 2316-65330.

[17] Gottardo, E.; Kaestner, C. A. A.; Noronha, R. V. "Estimativa de Desempenho Acadêmico de Estudantes: Análise da Aplicação de Técnicas de Mineração de Dados em Cursos a Distância". Revista Brasileira de Informática na Educação 22.01 p. 45. 2014.

[18] Rodrigues, R. L.; Medeiros, F. P. A. de; Gomes, A. S. "Modelo de Regressão Linear aplicado à previsão de desempenho de estudantes em ambiente de aprendizagem". Anais do Simpósio Brasileiro de Informática na Educação. Vol. 24. No. 1. 2013.

[19] Santos, R. N. dos; Siebra, C. A.; Oliveira, E. D. S. "Uma Abordagem Temporal para Identificação Precoce de Estudantes de Graduação a Distância com Risco de Evasão em um AVA utilizando Árvores de Decisão". Anais dos Workshops do Congresso Brasileiro de Informática na Educação. Vol. 3. No. 1. 2014.

[20] Rigo, S. J.; Barbosa, J.; Cambruzzi, W. "Educação em Engenharia e Mineração de Dados Educacionais: oportunidades para o tratamento da evasão". EaD & Tecnologias Digitais na Educação. V. 2.3. p. 30-40. 2014.

[21] Pequeno, H.; et al. "Uma Análise de Interação em Fóruns de EAD ".Anais do Simpósio Brasileiro de Informática na Educação. Vol. 25. No. 1. 2014.

[22] Oliveira Júnior, Roberto L. de; Esmin, A. A. "Monitoramento Automático de Mensagens de Fóruns de Discussão Usando Técnica de Classificação de Texto". Anais do Simpósio Brasileiro de Informática na Educação. Vol. 23. No. 1. 2012.

[23] Silva, L. A.; Trindade, D.; de Paula, C.; Pinto, S. et al. "Mineração de Dados em publicações de Fóruns de Discussões do Moodle como geração de Indicadores para aprimoramento da Gestão Educacional". Anais dos Workshops do Congresso Brasileiro de Informática na Educação. Vol. 4. No. 1. 2015.

[24] Silva, R. E. D.; Ramos, J. L. C.; Rodrigues, R. L.; Gomes, A. S.; Fonsêca, J. A. V. "Mineração de dados educacionais na análise das interações dos alunos em um Ambiente Virtual de Aprendizagem". Anais do Simpósio Brasileiro de Informática na Educação. Vol 26. No. 1. 2015.