Using demographics to understand better the students' behavior, improving the performance prediction in online courses

André R. Kuroswiski[1], Philip K. Chan[2]

[1]Department of Electronic Engineering - Instituto Tecnológico de Aeronáutica (ITA)
São José dos Campos, SP – Brazil

[2]School of Computing – Florida Institute of Technology (FIT)
Melbourne, FL – USA.

kuroswiskiark@fab.mil.br, pkc@cs.fit.edu

Abstract. *Predicting the students' final performance early in term, in online courses, can be an interesting resource to helping them changing their behaviors, aiming to improve their results. A feasible way to do this is applying data mining techniques in their activities' logs. Another important information that should help this analysis, is the students' demographics data, that give us important insights about the expected behavior of each student. However, previous works that tried to use the demographics to improve the performance prediction, did not get so good results. Our study proposes a different approach to use the demographics data, using only this information to generate clusters of students before applying the data mining techniques in the activities' logs, trying to group the students with more similar needs, increasing the chance to get better results. To evaluate the benefits of our method, we compared the accuracy gain in the prediction of the students' performance, using or not our approach. As results, using the demographics to generate cluster as the first step, and testing with four different methods commonly applied to this purpose, we had an average gain of 11.1% in the accuracy of the predictions.*

1. Introduction

The online courses are a reality for almost every student nowadays, and the utilization of Artificial Intelligence techniques to better understand how the students interact with the systems and how each behavior could affect the student's performance, become a very common and important field of study in Learning Analytics. A common target of many studies is to predict the students' final results using the system interaction logs, together or not, with the demographics information as features for classifiers algorithms. Our study presents an alternative and very simple method to improve the results of these predictions, using the students' demographics data to generate clusters before the application of classifiers algorithms to obtain the predictions.

The advantages of our approach are that, besides easily improve the predictions' accuracy for different algorithms, the analyze of the students' characteristics inside each cluster, showed that this method can help identifying more specific and important behaviors for each group of students, depending of their previous knowledge and known

DOI: 10.5753/cbie.sbie.2017.1577

characteristics, aiming to help them to improve their results, pointing some behaviors that they would change.

2. Related Work

The utilization of data mining techniques in the students' activity data in online courses to extract their behaviors characteristics and predict their performance is a common topic in Leaning Analytics, growing significantly last years. In [D'Inverno et al. 2016; Mori and Chan 2016; Ren et al. 2016], for example, we can see that the performance prediction is feasible and can provide satisfactory results through different methods. In [Graf et al. 2009; Lo et al. 2012], they used the activity data to identify different students' learning characteristics, showing that the interactions logs are a good source to understand the students' needs.

A common way to try improving the results is the utilization of more sources of data, for example, the demographics information about the students. However, the demographics data is not being commonly used in recent studies that try to predict the students' performance in online courses and, besides that, most concluded that this information would not help too much to identify the students' behaviors that affect their grades. For example, in [Brooks et al. 2015], they compared the difference in the prediction accuracy using or not the demographics data as features for a J48 classifier in Massive Open Online Course data, and concluded that there is no difference using this data. In [Wolff et al. 2013], they also added the demographics to the interaction data, trying to identify students at risk, but the advantages of using this data was practically insignificant. In [Lu et al. 2003], the demographics made part of a more complex model that try to determine the students' learning state, but, when they evaluated the model without demographics, the results were very similar. In [Guo and Reinecke 2014], together with learning styles and patters of learning, the demographics data did not show any significant correlation with the students' performance.

On other hand, despite many studies did not find correlation between demographics and students' performance in online courses, in the same study [Guo and Reinecke 2014], we can also see that, for many years, different studies have shown that the students' demographics affect directly the students' behavior in a large variety of courses. In a more recent study [Bouchet et al. 2013], they clearly identified different interactions patterns in online courses, depending of the students' demographics. Considering these findings, we should expect that the demographics data has information to help us to better identify the students' behavior characteristics and, consequently, would be possible to improve the performance of the predictions. Therefore, we looked up for a new way to use this data more efficiently, trying to extract more advantages from this significant information.

3. Approach

Since in the previous studies the demographics data were used alone, or mixed with another kind of data, and the results were not so good, our approach try to use the demographics data in a different way, instead of use it only as features for classification algorithms, we use this data to generate clusters of students as the first step, and after that, we apply, inside the clusters, the algorithms in the students' interaction data, generating the predictions. In this way, we expect to keep inside the clusters students that have more similar behaviors and needs to perform well in the course, and then, we would be able to

generate better predictions for the students' performance, since we would be applying classifiers in more similar students in those aspects.

With our method, besides the target of improving the prediction results, we also aim to improve the identification of important behaviors that the students could change to perform better , since, analyzing all the students together, the indicated features would not be too specific [Mori and Chan 2016], and then, not very helpful for this intent. Therefore, considering that the students inside the clusters would have more similar needs, the important behaviors identified could be more effective to help them to improve their studies.

In general, our approach is simple, but at the same time is very flexible, since can be easily applied in many other studies as the first step to improve the final results or to identify more specific characteristic of the students' behavior.

4. Experimental Evaluation

4.1. Dataset

To evaluate our approach, we used a dataset of a graduate online course called "Concepts and Principles of Behavior Analysis" that was offered from January 2013 to April 2013 at Florida Institute of Technology. From the course, we have information about the students' activities logs, grades, demographics data and the syllabus, from where we get the important dates and characteristics of the course. We have the data of 108 students in the course, those are very heterogeneous in terms of age, level of degree, field of degree and experience in the field, making this dataset very interesting to analyze the effects of these demographics in their behaviors.

4.2. Students' activities time estimation

An initial problem with the dataset is the uncertainty about the exact time that each student spent in each activity, mainly because many times the students did not performed the logout after they finish their studies. To reduce this problem, we created an estimator for the time spent in each activity by the student when the exact time is not available.

The estimator (Equation 1), calculate the proportion between the total time spent by the student in the course and the average time spent by all other students, then, the estimative is this proportion times the average time spent by all students in the activity that we need the prediction.

$$\tilde{t}_n(a_x) = \frac{T_n}{\overline{T_a}} \cdot \overline{t_a}(a_x) \qquad \text{(Equation 1)}$$

$t_n(a_x)$ → time estimated for student 'n' in the activity 'x'

T_n → total time spent by student 'n' in the course

$\overline{T_a}$ → average time spent by all students in the course

$\overline{t_a}(a_x)$ → average time spent by all students in the activity 'x'

Using this estimator, we obtained the time spent by the students in all activities that we did not had the exact values. Considering all students' activities, in the end, we

got that the estimated time represented 21% of all time stored in the data. Therefore, we should expect that the features evolving the time spent by students in specifics activities, will not be totally reliable.

4.3. General features

Using the data available, we generated a set of features to feed the classifiers algorithms, where some features were simply the brute data itself, the low-level features, while for others we merged or processed the data to generate the high-level features [Mori and Chan 2016]. For the low-level features, we extract the data direct from the students' interactions events, obtaining totals or averages time spent and the number of occurrences of specific activity. Table 1 presents the final low-level features selected for our tests.

Table 1. Low-level Features

Nº	Feature	Description	Nº	Feature	Description
1	Interactions	Number of clicks in using the system.	**8**	Videos times	Number of access to the videos material.
2	Total time	Total time spent using the system.	**9**	PPT times	Number of access to presentation materials.
3	Login times	Number of logins.	**10**	PPT time	Total time spent in presentation materials.
4	Logout times	Number of logouts.	**11**	Average tests time	Average time spent taking tests.
5	Guides times	Number of access to the study guide materials.	**12**	Discussions times	Number of access to the discussion forums.
6	Guides time	Total time spent reading guide times.	**13**	Supplemental times	Number of access to the supplemental materials.
7	Videos time	Total time spent watching videos.	**14**	Supplemental time	Total time spent in supplemental materials.

In addition, we generated eight high-level features, aiming to improve the utilization of all data available. Basically, the high-level features become from processed or merged data from different sources, for example, using the due dates from the syllabus, to measuring the students delayed or advanced tests' submission. Table 2 presents final high-level features selected for our tests.

Table 2. High-Level Features

Nº	Feature	Description
1	Time after units released	Average time the student took to read the units' materials after it has been released.
2	Time after videos released	Average time the student took to watch the units' videos after it has been released.
3	Tests before due	Average number of hours that the student took to take the tests before the due date.
4	Tests delays	Number of times that the student delivers the tests after the due date.
5	Study time before test	Average number of hours that the student spent studying during the last 3 days before the tests.
6	Study regularity	Index proportional to the concentration of the student's interactions events in the same days during the weeks.
7	Tests regularity	Index proportional to the concentration of the student's tests submission in the same day during the weeks.
8	Days off	Average number of days that the student did not studied during each week.

4.4. Clusters from Demographics

To implement our approach, the next step was using the demographics data to split the students in different clusters. To do it, we selected 4 main demographics aspects of the students as reference to perform the KMeans algorithm. As presented in the Table 3, we assigned values from 1 to 4 for the students' characteristics, in a such way that it would represent different influences that each aspect could have in the students' behavior during the course. Considering the results of [Bogarín et al. 2014] and [Bouchet et al. 2013], we decided to perform the tests with 2 and 3 clusters, trying to obtain the groups that keep the most similarities in the behaviors as possible.

Table 3. Demographics aspects selected to generate the clusters

N°	Feature	Value - Description
1	Age	[1] - 20 to 30 years [2] - 31 to 40 years [3] - 41 to 50 years [4] - 51 to 60 years
2	Highest Degree	[1] - Bachelors [2] - Post Graduate [3] - Masters [4] - Doctorate
3	Years of Experience in the Field	[1] - less than 1 year [2] - 1 to 4 years [3] - 5 to 8 years [4] - 8 or more years
4	Major of the Highest Degree	[1] - Psychology/Counseling/Child Development; [2] - Education; [3] - Speech/Language Pathology; [4] - Others;

4.5. Evaluation Criteria and Procedures

Considering our main goals, using the demographics to improve the performance prediction and the identification of the most important behaviors for the students' success, we evaluate our method in two ways:

- Compare the accuracy gain for predictions of the students' final results using or not the demographics to create the clusters before applying the classifiers: if the students grouped in the clusters have more similar needs and behaviors, we should expect better results in the prediction algorithms.

- Analyze how the most important identified behaviors for the students' performance changed after using the demographics to create the clusters: if we had a prediction's performance improvement and we identified differences in most important behaviors between the clusters, this difference indicate that the new groups have different needs.

4.6. Performance Results

To evaluate the performance of our method, we analyzed the prediction's accuracy gain through the utilization of the demographics to generate clusters in 4 different algorithms: C4.5, Random Forest, Artificial Neural Networks (ANN) and Naïve Bayes. What we tried to predict is if the final result of the student would be above or below the average. The prediction came from the high and low-level features generated for each student in a specific week in the middle of the course, using the cumulative data up to this date. Since

we aim to use these predictive results to help us to identify possible changes in the student's behavior, our tests went from the 3rd up to the 7th week in a ten weeks' course. We consider that, before the 3rd week we have too few data and, after the 7th, we do not have enough time to change the behaviors efficiently.

The evaluation considered a test set of 25% of all data, that was selected randomly, and the tests were repeated 50 times with different test sets to get an average result. A problem that we identified was that our dataset is not too big, so many times, when we tried to create 3 clusters, one of them were too small to apply all the classifiers. To solve it, we considered to stay only with two clusters, however, after some tests, we observed that 3 clusters could generate better results. The solution was to use a K-Nearest Neighbors algorithm, using N as 5, in the smallest cluster, and use the other classifiers in the bigger ones.

In this way, we obtained the results presented in the Table 4, that have the results for the four algorithms in different conditions. In the first column of results, the prediction considered only the features extracted from the logs. In the second, the demographics were merged to the data as extra features, while, in the third, the demographics were used to generate the clusters as the first step. Additionally, the accuracy gains are presented, representing how much the utilization of the demographics improved the results in both cases.

Table 4. Accuracy results and gains for the performances' prediction using the demographics as features or to generating clusters

	Accuracy			Accuracy Gain	
Demographics?	***No***	***Features***	***Clusters***	***Features***	***Clusters***
C4.5	0.60	0.61	0.69	0.5%	12.2%
Random Forest	0.67	0.67	0.73	0.2%	8.9%
Naive Bayes	0.64	0.64	0.73	-0.3%	12.1%
ANN	0.64	0.63	0.72	-1.6%	11.1%
Mean	0.64	0.64	0.72	-0.3%	11.1%

The results shown that the demographics itself, being used as features, did not help to improve the prediction performance, confirming the results of [Guo and Reinecke 2014; Lu et al. 2003; Wolff et al. 2013], with gains of 0.5% for C4.5 and 0.2% for Random Forest, and losses of 1.6% for the ANN and 0.3% for the Naïve Bayes. These results indicated that the demographics really may appear to do not be very useful when predicting students' results in online courses, however, using this data to generate the clusters of students as the first step, the gains were significant, reaching an average of 11.1% for all algorithms.

While this relative gain in the accuracy indicate that the method can be useful, the comparison of the absolute accuracy results with other solutions are not so direct. In [Shahiri et al. 2015] they made a review of many recent studies identifying huge variation in the performances prediction accuracies, from 50% to 98% and, beyond the variation of the algorithms applied in each one, the kind of data considered affected significantly the results. An analysis of the review indicated that the utilization or not of data that contain student partial grades or historical grades had a huge impact in the accuracy of the predictions, what could be clearly expected. For example, considering the 28 studies presented in the review, we have an average of 77.6% in the accuracy, however splitting

the studies between those used or not previous or partial grades as data, we have averages accuracies of 81.5% and 67.4% respectively.

In our study, since we aim to use the prediction to generate recommendations in the future, the utilization of partial grades as data did not show to be a good solution, because these results become the most important feature, suppressing the identification of specific behaviors that would be interesting to change. Indeed, adding the partial grades as features, we reached an average accuracy of 82.4%, however, in all these cases, the main possible recommendation extracted would be "improve your grades and you will have better results", clearly not too useful.

Regardless, the mean accuracy of our method is in-line with the recent studies, actually, without grades information, it is 4.6% above the average, considering the [Shahiri et al. 2015] review. Besides, the accuracy gain, despite [Brooks et al. 2015; Guo and Reinecke 2014; Lu et al. 2003; Wolff et al. 2013] results, confirmed that the demographics can be an interesting information to improve the performances' prediction in online courses.

4.7. Clusters Analysis

As we saw, using the demographics data allow us to split the students in groups with more similar behaviors, improving the prediction accuracy. Therefore, we could expect that, between the clusters, we can identify different characteristics that would help to better understand the students' needs.

First, we analyzed the remarkable demographics characteristics in each cluster, trying to identify some different main characteristics that could define the students in each group. In this case, the Clusters A and C were bigger with 34 and 62 students, and, in both, we were able to identify some main characteristics over all that can characterize the groups, however, in the Cluster B, with only 12 students, the characteristics were more variable, forming a more heterogeneous group of students. The results are in the Table 5.

Table 5. Remarkable demographics characteristics of each cluster

Feature	Cluster A	Cluster B	Cluster C
Age	< 30	All Ages	30-40
Highest Degree	Bachelor	1-Bachelor 2-Post-Degree	1-Master 2-Post-Degree
Experience	Lower	All Levels	Medium
Major	1-Psychology 2-Education	Other	1-Education 2-Psychology

After that, we analyzed how the characteristics in the students' behaviors affected their grades, and we found some interesting differences between the clusters. For this analyze we split the students of each previously defined clusters in two groups, based on their grades, through a Spectral Clustering technique, and then, we compared the average values of the features of all students in each group. For the Cluster B, the smallest and very heterogeneous, we did not find too distinguishes differences that could be useful, however, the Clusters A and C shown interesting results in some features as presented in the Figure 1. The values presented in the figure are the average of the normalized values of the students' features and, for each feature, we have the averages for the groups with higher and lower grades, inside each cluster.

Using the results to compare the Clusters A and C, we can identify different impact of features in the grades, for example, while in the Cluster A, the students with higher values in “Interactions”, “Videos times”, “Guides times” and “Study regularity” are the students with higher grades, in the Cluster C, we have the opposite, the students with lower grades had these behaviors.

At first, these results seem to be unexpected, since, in the Cluster C, the higher values in these features that indicate more study, did not represent better grades. However, the clusters’ remarkable characteristics identified before, Table 5, can help to understand it, since the Cluster C, in general, had students with more experience and higher levels of degree, so, not necessarily, they need to study too much to improve their performances. On other hand, the Cluster A, contain students with less experience and lower levels of degree, therefore, we can expect that they did not have previous knowledge about the topics, and then, the amount of study would be more directly related to their performances. Another features presented similar importance to the grades in both clusters, for example, higher values of “Tests before due” and “Discussion times” are directly related to higher grades in A and C, while higher values of “After units released” is directly related to lower grades.

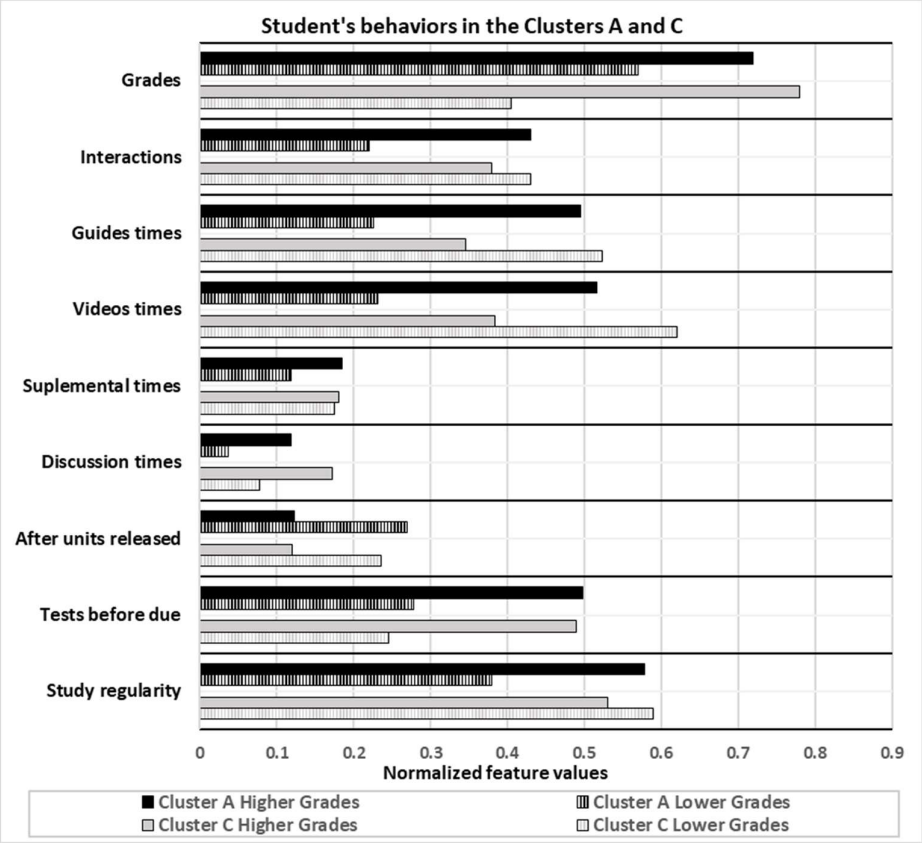

Figure 1 - Comparison of the students' behaviors in Clusters A and C, depending of their grades.

These results confirmed that splitting the students in clusters through the demographics can be an efficient way to identify, more specifically, important behaviors

for each group of students and, besides improving the predictions, allow the generation of more significant recommendations.

5. Conclusion

Differently of previous studies, our results demonstrate that the demographics data can help, not only to improve the accuracy of the students' performance predictions in online courses with an average gain of 11.1% in our tests, but also, can help us to understand better the students' behaviors, since the different needs to archive higher grades presented by the students in different clusters, indicate specific characteristics, depending of their demographics, which were hidden before.

Certainly, the utilization of only one curse dataset is a restriction that may affected the results, however, since significant gains in the predictions appeared consistently for different algorithms, and the differences in the students' behaviors become clear between the clusters, at least for the course studied, the results demonstrated that the method can be an interesting alternative to use the demographics data more efficiently. The size of the dataset also shown to be a problem, even more because we were splitting it, what become clear when using three clusters that, despite presented the best results, in one of them we had too few students and the group were significantly heterogeneous. Therefore, with a bigger dataset, we could have more clusters, splitting these remaining students in more homogeneous groups.

Finally, as the next step, we need to test the method with other datasets, preferably bigger and with more details about the students' interactions. However, at this point, the results confirmed that the idea can be useful and, mainly due the simplicity to apply the method, many other techniques can take advantages of this idea to use the demographics data to improve their performances.

References

Bogarín, A., Romero, C., Cerezo, R., & Sánchez-Santillán, M. (2014). Clustering for improving educational process mining. *Proceedins of the Fourth International Conference on Learning Analytics And Knowledge - LAK '14*, 11–15.

Bouchet, F., Harley, J. M., Trevors, G. J., & Azevedo, R. (2013). Clustering and Profiling Students According to their Interactions with an Intelligent Tutoring System Fostering Self-Regulated Learning. *JEDM - Journal of Educational Data Mining*, *5*(1), 104–146.

Brooks, C., Thompson, C., & Teasley, S. (2015). Who You Are or What You Do. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale - L@S '15* (pp. 245–248).

D'Inverno, M., Al-Rifaie, M. M., & Yee-king, M. (2016). Investigating Swarm Intelligence for Preformance Prediction. *Proceedings of the 9th International Conference on Educational Data Mining*, 264–269.

Graf, S., Kinshuk, & Liu, T.-C. (2009). Supporting Teachers in Identifying Students' Learning Styles in Learning Management Systems: An Automatic Student Modelling Approach. *Educational Technology & Society*, *12*(4), 3–14.

Guo, P. J., & Reinecke, K. (2014). Demographic differences in how students navigate

through MOOCs. In *Proceedings of the first ACM conference on Learning @ scale conference - L@S '14* (pp. 21–30).

Lo, J. J., Chan, Y. C., & Yeh, S. W. (2012). Designing an adaptive web-based learning system based on students' cognitive styles identified online. *Computers and Education*, *58*(1), 209–222.

Lu, J., Yu, C. S., & Liu, C. (2003). Learning style, learning patterns, and learning performance in a WebCT-based MIS course. *Information and Management*, *40*(6), 497–507.

Mori, M., & Chan, P. K. (2016). Identifying Student Behaviors Early in the Term for Improving Online Course Performance. *Proceedings of the 9th International Conference on Educational Data Mining*, 611–612.

Ren, Z., Rangwala, H., & Johri, A. (2016). Predicting Performance on MOOC Assessments using Multi-Regression Models. In *Proceedings of the 9th International Conference on Educational Data Mining (EDM)* (pp. 484–489).

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A review on predicting student's performance using data mining techniques. In *3rd Information Systems International Conference, 2015* (Vol. 72, pp. 414–422).

Wolff, A., Zdrahal, Z., Nikolov, A., & Pantucek, M. (2013). Improving retention: Predicting at-risk students by analysing clicking behaviour in a virtual learning environment. *LAK '13 Proceedings of the Third International Conference on Learning Analytics and Knowledge*, 145–149.