Unsupervised Learning for Understanding Student Achievement in a Distance Learning Setting

Shuangyan Liu and Mathieu d'Aquin
Knowledge Media Institute
The Open University, UK
{shuangyan.liu, mathieu.daquin}@open.ac.uk

Abstract—**Many factors could affect the achievement of students in distance learning settings. Internal factors such as age, gender, previous education level and engagement in online learning activities can play an important role in obtaining successful learning outcomes, as well as external factors such as regions where they come from and the learning environment that they can access. Identifying the relationships between student characteristics and distance learning outcomes is a central issue in learning analytics. This paper presents a study that applies unsupervised learning for identifying how demographic characteristics of students and their engagement in online learning activities can affect their learning achievement. We utilise the K-Prototypes clustering method to identify groups of students based on demographic characteristics and interactions with online learning environments, and also investigate the learning achievement of each group. Knowing these groups of students who have successful or poor learning outcomes can aid faculty for designing online courses that adapt to different students' needs. It can also assist students in selecting online courses that are appropriate to them.**

Keywords*—*unsupervised learning; cluster analysis; K-Prototypes algorithm; open learning analytics datasets; distance learning

I. Introduction

Research in learning analytics and education data mining concentrates on understanding learning that occurs in learning systems. Many studies have investigated the predictive factors of student success in online learning in particular student behaviours in the systems [1] [2] [3]. However, student characteristics such as demographic information are often neglected in the analysis of factors leading to successful or poor learning outcomes. The analysis of different factors that may influence students' learning achievement is hard because the data used for analysis are often mixed (numerical and categorical). It requires a complex learning analytic technique to analyse the mixed data.

The purpose of this paper is to address this challenge by applying unsupervised learning for investigating the influence of different student characteristics on learning achievement. We apply a clustering method on a distance learning course data set that is extracted from the Open University Learning Analytics dataset [4]. The objectives of our study consist of: *a*) group students based on their characteristics including demographic characteristics and their engagement in online learning activities, *b*) investigate the learning achievement of each group. The clustering method that we utilise, which is called the K-Prototypes clustering algorithm [13], is appropriate for working on data containing either numeric values or categorical values. We chose this method because most student demographic data is categorical (e.g. gender, previous education level, and region where they are living) and student behaviour data is numeric (e.g. number of times for viewing a learning forum). Our experiment by applying this approach to the collected test set shows an efficient way to find the common characteristics of students with same levels of achievement. We also show that what the common student characteristics are concerning different levels of learning achievement.

In the following sections, we describe the method to investigate the influence of different student characteristics on learning achievement in more detail. We then present the findings of a study on a distance learning course data set using this method.

II. Literature Review

A. Learning Analytics

Learning analytics is an emerging area that draws techniques from a number of communities such as artificial intelligence and data mining for understanding and improving learning [5] [6]. Greller and Drachsler provided a holistic view of the critical problems, the processes and requirements behind learning analytics in [7]. Chatti et al. proposed a reference model for classifying learning analytics research and also identified challenges and opportunities in this area [6].

Ratnapala and Deegalla [1] utilised the K-Means clustering method [8] on data collected from two e-learning engineering courses for analysing patterns of students' access behaviour. Lee et al. presented a visual data analytic method [2] to understand how patterns of student interaction with a learning management system are related to their learning outcome. In addition, Akçapnar et al. [3] applied the Self Organising Map clustering method [9] to identify distinct groups of students by their interaction with an online learning environment.

B. Cluster Analysis

Cluster analysis is one type of unsupervised learning algorithms. The task of cluster analysis or clustering is classifying objects in homogeneous groups (called a cluster). Objects in the same cluster are more similar to each other than objects in other clusters. Kaufman and Rousseeuw provided a systematic view of the most practical cluster analysis methods in [10].

Centroid-based clustering algorithms such as the K-Means algorithm are the most widely used clustering algorithms. The general idea of centroid-based clustering is that clusters of objects are represented by a central vector that is usually made up of means or modes of the feature values of objects in the cluster. The task of K-Means algorithm is to find k centroids of the data set and assign objects in the data set to the nearest centroid, such that the squared distances from the centroids are minimised. More information about the K-Means algorithm can be found in [8]. The drawback of K-Means algorithm is that it works only on numeric values, which prohibits it from being used for clustering practical data set containing categorical data.

C. K-Prototypes Algorithm

The K-Prototypes algorithm is an extension to the K-Means algorithm for clustering objects described by mixed numeric and categorical attributes [11]. It integrates the K-Means and K-Modes processes [11] to cluster data with mixed attributes. To be more concrete, K-Prototypes applies the squared Euclidean distance measure on numeric attributes and the simple matching dissimilarity measure [10] on categorical attributes for finding the closest centroids of objects.

The K-Prototypes clustering process is similar to the K-Means process except that it uses the K-Modes approach to updating the categorical values of cluster centroids or prototypes. Thus K-Prototypes preserves the efficiency of the K-Means algorithm. In this paper, we apply the K-Prototypes algorithm for the clustering process since we aim to use both student demographic data and their interactions with a virtual learning environment system, which is comprised of both numeric and categorical attributes.

III. Problem Statement

Regarding the clustering process to be discussed below, we only consider two general data types, numeric and categorical. According to the K-Prototypes algorithm [11], we provide the following definitions to describe the problem that we want to address in this paper.

Definition 1. A student object X encompasses r numeric attributes and m categorical attributes, which is represented using a vector
$[x_1, x_2, ..., x_r, x_{r+1}, x_{r+2}, ..., x_m]$.

Let $\boldsymbol{X} = \{X_1, X_2, ..., X_n\}$ be a set of n student object. A student object X_i can be represented as $[x_{i,1}, x_{i,2}, ..., x_{i,m}]$. According to the simple matching dissimilarity measure [10], the dissimilarity in categorical attributes between two student objects X, Y is represented as

$$dissim(X, Y) = \sum_{j=r+1}^{m} \delta(x_j, y_j) \tag{1}$$

where

$$\delta(x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases} \tag{2}$$

Let k be the number of clusters to be formed, then we have the following definition of the centroids of the k clusters.

Definition 2. The centroids of the k clusters is a set of objects $\boldsymbol{Q} = \{Q_1, Q_2, ..., Q_k\}$ that minimises the cost function

$$J = \sum_{l=1}^{k} \left(\sum_{i=1}^{n} \sum_{j=1}^{r} (x_{i,j} - q_{l,j})^2 + \gamma \sum_{i=1}^{n} \sum_{j=r1}^{m} \delta(x_{i,j}, q_{l,j}) \right) \tag{3}$$

where $\boldsymbol{Q}$ is in the same domain of $\boldsymbol{X}$ but Q_l is not necessarily a member of $\boldsymbol{X}$, and Q_l is represented as $[q_{l,1}, q_{l,2}, ..., q_{l,m}]$; the weight γ is used to avoid favouring either type of attribute.

In summary, the goal of our clustering process is to find the centroids or prototypes of the k clusters which minimise the cost function and partition a set of student objects into k clusters with each object belonging to the nearest cluster centroid. Moreover, we also want to investigate the learning achievement of each cluster and the common characteristics of students which have obtained the same levels of learning result.

IV. Approach and Data Set

We built a data set which we call Open Learning data set for our experiment, which is available online at the link given below[1]. It contains data about a distance learning course that was extracted from the Open University Learning Analytics dataset[2]. Aspects of the Open Learning data set includes demographic information of students who registered for the course, student interactions with the virtual learning environment for the course, and final results of students.

The data preparation process involved selecting, joining, encoding and cleaning data from the source dataset. To simplify the data structure for the clustering process, we created a table to contain all the aspects of data that we want to use. Each row of the table represents a student object identified by a unique key, i.e. id of the student. The columns of the table stand for the categorical or numeric attributes of a student. The last column of the table represents the final results of students who have taken the course. In the following subsections, we describe the different attributes of a student object and the encoding rules for converting the categorical data into a format that benefits the clustering process.

A. Categorical Attributes

There are six categorical attributes of a student object, which includes:

- Gender: the student's gender which takes the value of male (1) or female (2).
- Region: the geographic region where the student lived while taking the course. There are 13 regions in the data set. They were labeled from 1 to 13. Details of the encoding mapping can be found at the description page of the data set.
- Highest Education: the highest education level that a student obtained on entry to the course. Labels 1 to 4 stand for "lower than A level", "A level or equivalent", "HE qualification", and "Postgraduate qualification" respectively.

[1] https://github.com/jennyindcs/Open-Learning

[2] https://analyse.kmi.open.ac.uk/open_dataset

- Index of Multiple Deprivation (IMD) Band: represents the level of deprivation for the place where the student lived during the course. It ranges from 0-10% to 90-100% representing the most deprived places to the least deprived places. We used numbers 1 to 10 to represent the ten bands individually.
- Age Band: band of the student's age which takes three values, "0-35", "35-55", "$\geq$55". They were encoded as the numbers 1 to 3.
- Disability: indicates whether the student has declared a disability (1 for "Yes", 2 for "No").

B. Numeric Attributes

Three numeric attributes are presented in the data set. They include:

- Previous Attempts: the number of times that a student has attempted this course.
- Studied Credits: the total number of credits for the courses that the student is currently studying.
- Sum of Clicks: the number of times the student interacts with the course material for the duration of the course. Students' interactions with different types of material (e.g. visits a course content website or course's assessment resources) were added up to obtain the values for this attribute.

The Open Learning data set contains information about 748 students who have taken the course at two different terms. Among them, 36 students have taken the course twice as they have failed or withdrawn from the course for the first time. To keep the data set consistent, data of those students from the second term were removed.

Another problem with the original data is that there are several missing values in the attributes of "index of multiple deprivation band" and "sum of clicks". Since the K-Prototypes algorithm for the clustering process does not accept missing values in numerical values, we have filled the missing values with zeros for the numeric attribute "sum of clicks". For the categorical attribute, we used zeros to represent the missing values as an "unknown values" category so that the algorithm treats unseen data as matching with each other but mismatching with non-missing data when determining the similarity between points.

V. Experiments

This section introduces the platform, implementation, and configuration that were applied for the K-Prototypes clustering process. Then it presents the results of the experiments, including an appropriate value of the parameter K, the partition of the student data, the discovered connections between student characteristics and their learning outcome.

A. Experimental Setting

We have implemented the K-Prototypes algorithm [11] using Octave [12] programming language[3]. We built our implementation of the K-Prototypes algorithm for two reasons. First, we had not found a publicly available implementation of K-Prototypes in other popular programming languages (e.g. MATLAB, R, WEKA/Java) for machine learning except Python to the best of our knowledge when the experiments were performed. Second, compared with Python, Octave has a light syntax which is more appropriate for fast prototyping.

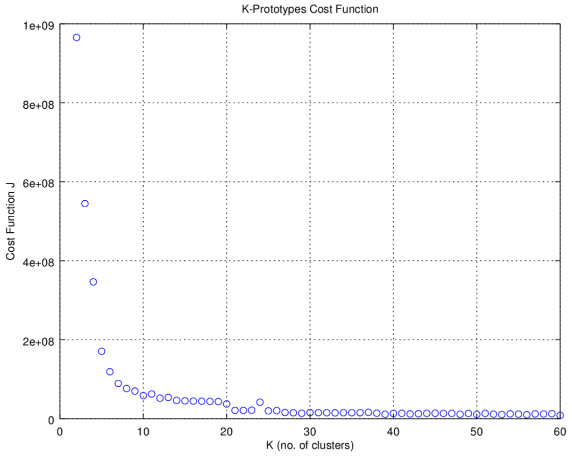

Fig. 1. Distribution of the cost function (3) for different K values (K=2 to 60)

Our implementation of the K-Prototypes algorithm sets the weight γ in (3) to be $0.5 * Xnum.std()$ by default [13] if the value of γ is not provided by the user. The $Xnum$ symbol stands for a matrix of the numeric values of a data set; $std()$ represents for the standard deviation of the elements of the matrix.

The results of K-Prototypes clustering are influenced by two important aspects, initialisation of the centroids and selection of the number of clusters (denoted by K). To avoid local optimal solutions, we utilised a random initialisation method and executed the algorithm for some times with different centroid seeds. Additionally, we applied the so-called "Elbow" method to determine which value of K to use. That is, we carried out the experiment with different K values (from 2 to 60) and picked the one from which the value of the cost function decreased the most. The results will be discussed in the following subsection.

B. Results

The distribution of the cost function values per the number of K values for the K-Prototypes clustering process is shown in Fig. 1. The cost function values when K=2 is high, and it is getting lower when K is increased. The distortion goes down rapidly from 2 to 7, and then the distortion goes down very slowly after that. Therefore, the results obtained for K=7 have been considered for the evaluation.

Fig. 2 shows the distribution of instances for the clusters that were obtained after the clustering process (K=7). The centroids or prototypes, of each cluster are presented in Table I. The categorical attributes of the centroids are represented using the encoding of these attributes (which have been described in Section IV-A).

It is interesting to see that the largest cluster (K=5) has the lowest number of times for interacting with the course material

[3] https://github.com/jennyindcs/K-Prototypes

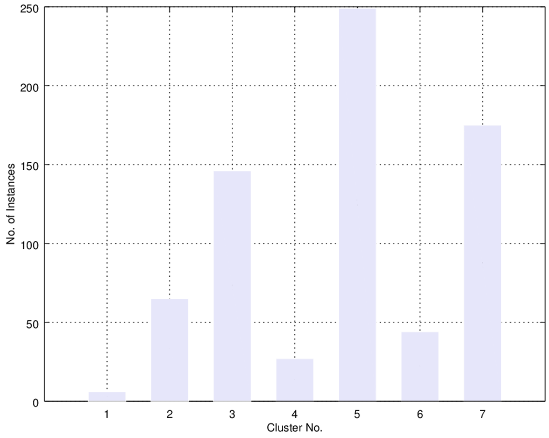

Fig. 2. Instances distribution of clusters

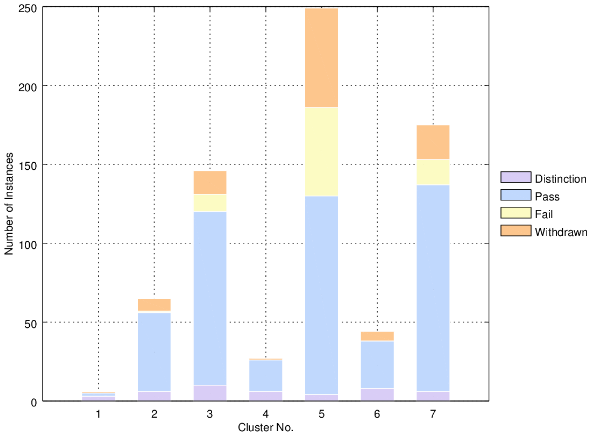

Fig. 3. Student final result distribution of clusters

TABLE I. CENTROIDS OF CLUSTERS

Attributes	Centroid Value						
	Cluster No.						
	1	2	3	4	5	6	7
Gender	1	1	1	1	1	2	1
Region	9	7	1	10	9	9	1
Highest education	2	2	2	3	2	2	2
IMD band	10	8	8	10	10	9	10
Age band	2	2	2	2	1	2	1
Disability	2	2	2	2	2	2	2
Previous attempts	0	0	0.007	0	0.008	0	0.011
Studied credits	60	73	76	76	97	67	85
Sum of clicks	12964	3064	1988	6733	396	4474	1156

(sum of clicks = 396), and the smallest cluster (K=1) has the highest number of clicks (sum of clicks = 12964). Comparing the two groups, they are different in the students' age and the total credits for the courses that the students were studying. Cluster 1 is comprised of older students (50% 35-55 years old and 50% $\geq$55 years old) while Cluster 5 consists of younger students (64% 0-35 years old, 35% 35-55 years old and 1% $\geq$55 years old). Cluster 1 and cluster 5 have the lowest and highest credits respectively. It indicates that older students who were studying for fewer courses tended to interact more with the virtual learning environment than younger students who were taking more courses at the same time. Additionally, the average credit for the whole class of students is 84. Only two clusters (K=5 and K=7) are above this average value.

We computed the number of instances for different levels of students' final result for the obtained clusters (Fig. 3). The final results of the students for the course are classified into four levels: distinction, pass, fail and withdrawn. As can be seen from Fig. 3, cluster 1 and cluster 4 contain students that have achieved only successful learning outcome (distinction or pass final result). The students in these two clusters were made up of mainly male students who had an age of 35-55 with no disability. 36% of them came from the south and southwest regions of UK, and most of the students came from the middle to the most privileged living places. Most of the students in cluster 1 and cluster 4 had obtained A level and higher education qualifications as the highest education level. In addition, these students had not attempted the course before and were studying courses with credits lower than the average value (84) of the whole class, but they visited the course material most frequently compared with other clusters.

Cluster 5 has the largest portion of students (nearly half) who have obtained unsuccessful learning outcome (fail or withdrawn). The centroid of cluster 5 shows that this cluster is a relatively younger group ("age band"=1) comprised of passive students who had the lowest number of times for accessing the virtual learning environment ("sum of clicks"=396). It also shows that the students in this cluster were studying for multiple courses at the same time during the investigated distance learning course ("studied credits"=97 which is the highest average value compared to other clusters). This may be a factor of why the students had the lowest total number of interactions with the learning environment.

We have found that three clusters (cluster 2, 3, 7) are mainly made up of students who have passed the course. The common student characteristics that these clusters possess include that they contain mainly male students who have only received A level education experience before. They are alike in the total number of interactions with the virtual learning environment. The students in the three clusters were less passive in terms of interactions with the learning environment since they completed a middle level of the sum of interactions with the learning environment.

VI. Conclusion

In this paper, we presented a study of applying a clustering method to investigate the influence of different student characteristics on student's learning achievement. The contributions of this paper include: *i*) we presented an approach to identify the influence of student characteristics on learning achievement using the K-Prototypes cluster analysis algorithm; *ii*) we built a data set that contains mixed data of categorical and numeric values about student characteristics and learning activity information for a distance learning course; *iii*) we implemented the K-Prototypes algorithm using the Octave language which is publicly accessible online; *iv*) a set of experiments using the K-Prototypes algorithm with different K values was performed, and our experiment shows the potential of our approach to finding the common characteristics of students with same levels of achievement. Our approach based on the K-Prototypes clustering algorithm can be efficiently applied to large data sets.

Our study shows some interesting findings of the relationship between student characteristics including their interactions with the virtual learning environment and learning achievement. Comparing groups of students who obtained successful and unsuccessful learning outcomes, the results showed that "successful" groups encompassed more mature active students than "unsuccessful" groups. It also suggested that "successful" groups contained a larger percentage of students who were living in the most privileged areas than the "unsuccessful" groups. We also found that "successful" groups consisted of a larger percentage of students who have obtained higher education levels than "unsuccessful" groups.

References

[1] I. Ratnapala, R. Ragel, and S. Deegalla, "Students behavioural analysis in an online learning environment using data mining," in *Proceedings of the 7th IEEE International Conference on Information and Automation for Sustainability*, 2014, pp. 1–7.

[2] J. E. Lee, M. Recker, A. J. Bowers, and M. Yuan, "Hierarchical cluster analysis heatmaps and pattern analysis: An approach for visualizing learning management system interaction data," in *Proceedings of the 9th International Conference on Educational Data Mining*, 2016.

[3] G. Akçapỳnar, A. Altun, and E. Cosgun, "Investigating students' interaction profile in an online learning environment with clustering," in *Proceedings of the 14th IEEE International Conference on Advanced Learning Technologies*, 2014, pp. 109–111.

[4] J. Kuzilek, M. Hlosta, D. Herrmannova, Z. Zdrahal, and A. Wolff, "Ou analyse: analysing at-risk students at the open university," *Learning Analytics Review*, pp. 1–16, 2015.

[5] G. Siemens. (2010) What are learning analytics? [Online]. Available: http://www.elearnspace.org/blog/2010/08/25/what-are-learning-analytics/

[6] M. A. Chatti, A. L. Dyckhoff, U. Schroeder, and H. Thüs, "A reference model for learning analytics," *International Journal of Technology Enhanced Learning*, vol. 4, no. 5-6, pp. 318–331, 2012.

[7] W. Greller and H. Drachsler, "Translating learning into numbers: A generic framework for learning analytics." *Educational technology & society*, vol. 15, no. 3, pp. 42–57, 2012.

[8] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[9] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[10] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009, vol. 344.

[11] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data mining and knowledge discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[12] J. W. Eaton, D. Bateman, and S. Hauberg, *Gnu Octave*. Network Thoery London, 1997.

[13] Z. Huang, "Clustering large data sets with mixed numeric and categorical values," in *Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining,(PAKDD)*. Singapore, 1997, pp. 21–34.