2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)
March 17–18, 2021, Amity University Dubai, UAE

2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) | 978-1-6654-2921-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCIKE51210.2021.9410741

Student Interest Group Prediction using Clustering Analysis: An EDM approach

Vedant Bahel
Department of Information Technology
G H Raisoni College of Engineering
Nagpur, India - 440016
vbahel@ieee.org

Shreyas Malewar
Department of Comp. Science & Engg.
G H Raisoni College of Engineering
Nagpur, India - 440016
shreyasmalewar@gmail.com

Achamma Thomas
Department of Artificial Intelligence
G H Raisoni College of Engineering
Nagpur, India - 440016
achamma.thomas@raisoni.net

***Abstract*—This paper proposes a clustering-based approach to identify and predict a suitable interest group for students in higher education system. Student interest group stands for on-campus students club that reflects the co-curricular or extra-curricular participation of students apart from general academics. Such interest groups play a vital role in development of a student's overall personality. K-means clustering algorithm has been used for this purpose. The experiment has been carried out on data collected by surveying students in higher education space. The purpose of this survey is to capture interest features of the student which is fed to the clustering algorithm. The overall concept ensures streamlining of the student's efforts to maximize success.**

***Keywords*—Domain Modelling, Educational Data Mining, Computational Intelligence and Clustering**

I. Introduction

Data Science has become a promising field of research delivering commendable applications in every domain including education. The field of education was relatively less explored by the researchers. However, lately the area of research popularly known as "Educational Data Mining" (EDM) has been widely studied. The concept of EDM aims to solve major problems in educational space i.e. higher education as well as K-12 education, using the concept of statistics, data mining and machine learning. Another term used for this field of research is called "learning analytics". More likely this field of research discusses the student's learning habits in an online learning environment or analysis of student's data for decision making. Such a concept is very useful for decision making from the perspective of the academic council and policymakers. Students are often benefited by the systems and applications derived from these researches. In [1], authors have discussed various applications of computational intelligence in the higher education space. Authors have discussed some of the widely used applications like GPA prediction, course suggestion module and decision taking on feedback-based inferences. In [6], authors have used EDM to track student performances and typical progression about the same in order to warn and provide early support to low performing students. One such application proposed in [1] is "Domain Modelling". Here, "Domain" refers to on-campus clubs or interest groups such as technical clubs, speaking clubs, dance clubs, etc. Most of the institutes or universities have such on-campus students' clubs where each club has a different set of aims and objectives. Students often find it difficult to choose a club they wish to be a part of. Domain Modelling, as a whole, refers to a prediction model that suggests the best suited Interest Group (IG) for a student based on certain inputs from the student. In this paper, the concept is named as" Student Interest Group (SIG) Prediction" instead of the previously used term in "Domain Modelling" [1].

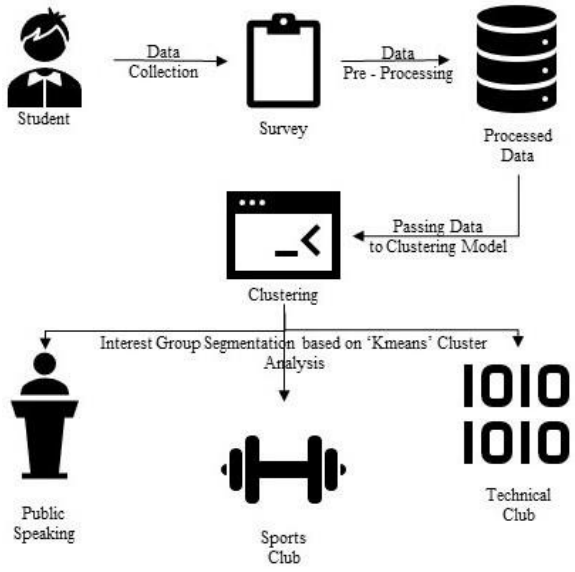

Fig. 1. General flow of the proposed model

As illustrated in "Fig. 1", in this paper, three interest groups are considered, technical club, communication club and sports club. A survey was carried out to record the latent skill set of students based on certain questionnaires. The corresponding club of that student was also recorded in the survey as a ground label for that data point. Later, statistical tools were implemented to analyze and understand the survey results after preprocessing the data. Clustering algorithm was used to model the data and to divide the data point into three clusters. Later, the clustering results were compared with the original label for all the data points in the data. Although the

978-1-6654-2921-4/21/$31.00 ©2021 IEEE

data had labels for all data point, an unsupervised system was considered. This was done to justify that SIGs can be identified even in a new space for new students. Thus, clustering which is an unsupervised concept was used.

II. Background

In [2], Zukhri et. al. has proposed a model which aims at clustering or grouping out students to form specific classes or batches. The aim of this is to ensure diversity in each class and to maintain uniformity in classes. The concept used here is based on Genetic Algorithm (GA). Similarly, in [3] the author proposes a sorting-based model for student's composition for teacher allocation depending on certain factors like minority, poverty rate, etc. For adapting educational strategies, it is very important to study student's typologies. In [4], authors have implemented an unsupervised machine learning based model to extract student's typologies. K-Means algorithm used in this paper, is a popular clustering algorithm commonly used in the areas of customer segmentation, bioinformatics, social network analysis, recommendation system, etc [7].

TABLE I
USE OF CLUSTERING ALGORITHMS IN EDM RESEARCH SPACE

Application	*Type of Clustering*	*Source*
Classification via clustering for predicting final marks based on student in forums	Comparative analysis (EM, Farthest First, Hierarchical Cluster, Simple K-means, x-Means, xIB). Best accuracy k-Means	[9]
Predicting students' final performance from participation in on-line discussion forums	Hierarchical Cluster, K-means, EM, Farthest. First. Best accuracy – k-Means	[10]
Reducing Dropout Ratio (An approach of improving student academic performance by using k means clustering algorithm and Decision tree)	K-means	[11]
Profiling individual differences in student motivation: A longitudinal cluster-analytic study in different academic contexts	-	[12]
Cluster Analysis in Higher Education Research	K-means	[13]
Improving Student's Performance Using Data Clustering and Neural Networks in Foreign-Language Based Higher Education	K-means	[14]
Student Performance Analysis Using Clustering Algorithm	K-means	[15]
Monitoring Student Performance Using Data Clustering and Predictive Modelling	K-means	[16]

In [8], Durairaj et. al. used K-Means clustering approach along with decision tree and Naive Bayes algorithm to analyze and predict students' performance. The model tries to segment students into different performance groups that can be further used to take necessary actions and decisions to improve performance of the student. Such applications have also been defined in the [1], however mostly students used the students' academic performance data to predict further performance unlike segmenting them as demonstrated in [8]. "Table. I" represented significant contributions in the EDM which uses clustering algorithms. Although there is no significant work done earlier, related to the title of this paper. However, K-Means clustering is proving to be widely accepted model for student analysis in the EDM domain.

III. Survey

In order to yield semantic results, well defined data is important. The data used in this research was collected by a survey. An important aim of this survey was to capture the features that accurately describe the interest of students in specific clubs or IG. The survey also captured the ground label for every data point so that it can be used to compare the results that were obtained after cluster modelling.

The survey was limited to students in the author's organization so as to maintain uniformity in the data. The data consists of a total of 176 data points. "Table II" shows the questions that were collected in the survey along with the corresponding attributes for each question. Throughout the paper, 'TC' refers to a technical club, 'CC' refers to a communication club and 'SC' refers to a sports club.

TABLE II
FEATURES & ITS CORRESPONDING ATTRIBUTES CAPTURED BY THE SURVEY

Question	*Affect Labels*	*Skill Captured*	*Response Type*
1	All	GPA	Linear (Out off 10)
2	TC	techSkill	Categorical (1-5)
3	TC & CC	presentationSkill	Categorical (1-5)
4	TC	codingSkill	Categorical (1-5)
5	TC & CC	documentationSkill	Categorical (1-5)
6	TC & CC	emailSkill	Categorical (1-5)
7	All	confidence	Categorical (1-5)
8	SC	affectFailure	Categorical (1-5)
9	SC	competitiveness	Categorical (1-5)
10	SC	notGoalOriented	Categorical (1-5)
11	SC	playSports	Binary
12	TC	techProject	Binary
13	SC	exercise	Binary
14	SC	stamina	Categorical (1-5)
15	CC	summit	Categorical (1-5)
16	CC	publicSpeaking	Binary

Before modelling, pre-processing was performed on the data by normalizing the GPA column that re-scaled the data between 0 and 1. All the other columns were declared as categorical input data.

IV. Implementation & Results

"Fig. 2" is the correlation matrix of the input features. As it can be seen some of the input features are very much related with each other. Most of them are corresponding to the "Sports Club" as target feature, thus assuring to get promising

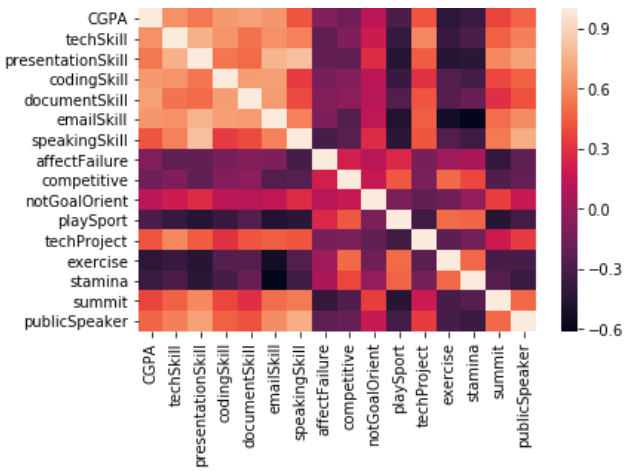

Fig. 2. Correlation heat map matrix between all input features

results. The matrix in some way also provides information about the significance of presence of the feature. For example, the features *"notGoalOriented"* does not show a good relationship with the other features which provides inference that it would not support dividing the data points into individual clusters.

Authors implemented a clustering model to fit the data and cluster the data points. Clustering is an unsupervised machine learning approach that does not require target labels. Authors implemented K-Means clustering algorithm [5]. K-Mean clustering adapts better to a new data space. It also serves as an advantage due to its easy scaling feature for massive dataset. K-Means clustering algorithm defines 'k' defined cluster centers based in the data space. All the data points are thus clustered around a cluster that has the nearest distance from the cluster center to that point. The value of 'k' (here 3) is defined by the user.

After K-Means clustering algorithm was carried out, the results obtained are shown in "Table. III". The clustering model was "*fit_intercepted*" on the training data itself to get the predicted label. These labels were then compared with the ground label that was collected via survey.

TABLE III
ACCURACY OF THE PREDICTED LABEL WITH GROUND LABEL

Output Label	***Accuracy***
Technical Club	76.19%
Communication Club	93.22%
Sports Club	88.67%
Overall	86.02%

In order to confirm that the cluster obtained belongs to the corresponding SIG as expected, the representative point of original data was compared to the centers of the obtained clusters. The representative point refers to the average of the data point of the individual labelled data group. The points obtained justified the obtained results.

A. Technical Club

This cluster consists of 42.28% of the total data points. It has been observed that this cluster of students have an average GPA of 9.12 which is significantly greater than the average of the other two clusters. Almost 60% of students in this cluster don't exercise daily as per the survey conducted. 78% of students in this cluster, sometime or the other have been involved in technical projects.

B. Communication Club

This cluster consists of 26.28% of the total population in the study space. The average GPA of students in this cluster is 8.77 which is significantly lesser than that of the technical club. Most of the students in this cluster have relatively better documentation and presentation skills. Similar to "Technical Club", "Communication Club" has 60% of students that do not exercise. Overall, it was observed that though "Communication Club" stands as an individual cluster, but the parameters stand very closer to the technical club. This also is a drawback of the model.

C. Sports Club

This cluster consists of 31.42% of the total population in data space. The average GPA of this cluster is considerably low i.e. 7.86. Features value of this cluster mainly, *exercise, stamina, GPA* makes the cluster highly segmented in the data space, thus clearly identifying the appropriate data points. This justifies why the accuracy for this cluster is higher than the other two.

These inferences made it important to closely observe the clusters and their features. It could be understood from the study that "Sports Club" definitely stands as an individual cluster with unique sets of values. However, the point of concern remains with the "Technical Club" and "Communication Club". Authors in "Table. IV", used a statistical tool namely "t-test". t-test is an inferential statistical method that checks whether the averages of two groups of features are reliably different or not. The reliability of the differences is calculated by taking in consideration the variance of the features and is represented by "*t-value*". The corresponding "*p-value*" represents the probability that the results can be produced by random trials. This means, lesser is the p-value, reliable is the model. The general threshold considered for reliable p-value is 0.05. This helps to generalize the finding beyond the population to get true analysis of the model.

As it can be observed in "Table. IV", features namely, *GPA, techSkill, presentationSkill, emailSkill, confidence, affectFailure, competitiveness, stamina summit* showed significantly lesser *'p-value'* which proves that the difference their value in both the clusters are significant. Whereas some of other features have extremely higher *'p-value'* which reflects that such mean differences can even be produced by randomised trials and are highly un-significant. Surprisingly, *codingSkill* which intuitively inclines more towards technical club is also a non-significant feature. The reason might be because not every person who is interested in technology knows coding. Overall, the t-test results provide an insight to the developer to define a better survey that can be used to better define SIG.

TABLE IV

FEATURES & T-TEST VALUE OF THE STUDENTS IN TECHNICAL (N=74) & COMMUNICATION CLUB (N=46) TO ANALYSE SIMILARITY/DIFFERENCES OF THE CLUSTER

Feature (df=118)	*Cluster*	*Means*	*Std. Dev.*	*t-value*
GPA	CC	8.77	0.397	-4.556
	TC	9.12	0.409	(p ≤ 0.01)
techSkill	CC	3.80	0.991	-4.556
	TC	4.51	0.575	(p ≤ 0.01)
presentationSkill	CC	2.08	0.577	-15.62
	TC	4.21	0.658	(p ≤ 0.01)
codingSkill	CC	4.28	0.746	1.012
	TC	2.08	0.702	(p = 0.313)
documentationSkill	CC	3.10	0.633	-7.470
	TC	4.13	0.776	(p ≤ 0.01)
emailSkill	CC	3.86	0.849	-2.540
	TC	4.20	0.544	(p ≤ 0.01)
confidence	CC	4.19	0.850	4.498
	TC	3.39	0.997	(p ≤ 0.01)
affectFailure	CC	2.30	0.620	-2.594
	TC	2.82	1.255	(p ≤ 0.01)
competitiveness	CC	3.71	0.742	-2.88
	TC	4.16	0.854	(p ≤ 0.01)
notGoalOrientednSkill	CC	2.65	0.476	1.60
	TC	2.41	0.885	(p = 0.102)
playSports	CC	0.45	0.498	-0.45
	TC	0.5	0.5	(p = 0.6)
techProject	CC	0.71	0.450	-0.82
	TC	0.78	0.411	(p = 0.413)
exercise	CC	0.391	0.488	-0.15
	TC	0.40	0.490	(p = 0.879)
stamina	CC	3.21	0.998	-3.152
	TC	3.72	0.758	(p ≤ 0.01)
summit	CC	4.43	0.770	2.866
	TC	4.01	0.779	(p ≤ 0.01)
publicSpeaking	CC	0.73	0.439	-0.0496
	TC	0.74	0.436	(p = 0.96)

V. CONCLUSION

The results obtained assures that the extended concept of the proposed system can be used to design a system at university which clusters students into different groups based on their skills and interest captured via a simple form. Such systems will be highly productive and will serve as a computer-based guidance mechanism for students. Some of the key points to be looked upon is improving the quality of survey such that it records in natural characteristics of students and forms a better cluster, which can be obtained by better model parametrization of features. In future, the aim is to design a full suite architecture that considers student's information which includes: academic history, skill-set, exam scores, communication skill, employ- ability skills, attendance summary and socio-economic status to determine the students habit and recommend measures to improve the overall profile. The system will aim to generate students reports and will provide combined student performance including all the applications mentioned in [1]. Additionally, the system will also serve as a recommendation engine for students' course, subject, and post- graduate course suggestion for the UG students.

REFERENCES

[1] Bahel, Vedant, Preeti Bajaj, and A. Thomas. "Knowledge Discovery in Educational Databases in Indian Educational System: A Case Study of GHRCE, Nagpur." In 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), pp. 235-239. IEEE, 2019.

[2] Zukhri, Zainudin, and Khairuddin Omar. "Implementation of genetic algorithms to cluster new students into their classes." Jurnal Fakultas Hukum UII (2006).

[3] Bonesrønning, Hans, Torberg Falch, and Bjarne Strøm. "Teacher sorting, teacher quality, and student composition." European Economic Review 49, no. 2 (2005): 457-483.

[4] Najdi, Lotfi, and Brahim Er-Raha. "Implementing cluster analysis tool for the identification of students typologies." In 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt), pp. 575-580. IEEE, 2016.

[5] Pedregosa, Fabian, Ga ël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel et al. "Scikit- learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.

[6] Asif, Raheela, Agathe Merceron, Syed Abbas Ali, and Najmi Ghani Haider. "Analyzing undergraduate students' performance using educational data mining." Computers & Education 113 (2017): 177-194.

[7] Kodinariya, Trupti M., and Prashant R. Makwana. "Review on determin- ing number of Cluster in K-Means Clustering." International Journal 1, no. 6 (2013): 90-95.

[8] Durairaj, M., and C. Vijitha. "Educational data mining for prediction of student performance using clustering algorithms." International Journal of Computer Science and Information Technologies 5, no. 4 (2014): 5987-5991.

[9] Lopez, Manuel Ignacio, Jose Maria Luna, Cristobal Romero, and Sebastian Ventura. "Classification via clustering for predicting final marks based on student participation in forums." International Educational Data Mining Society (2012).

[10] Romero, Crist óbal, Manuel-Ignacio López, Jose-María Luna, and Se- bastián Ventura. "Predicting students' final performance from participa- tion in on-line discussion forums." Computers Education 68 (2013): 458-472.

[11] Shovon, Md, Hedayetul Islam, and Mahfuza Haque. "An Approach of Improving Students Academic Performance by using k means clustering algorithm and Decision tree." arXiv preprint arXiv:1211.6340 (2012).

[12] Br åten, Ivar, and Bodil S. Olaussen. "Profiling individual differences in student motivation: A longitudinal cluster-analytic study in different academic contexts." Contemporary Educational Psychology 30, no. 3(2005): 359-396.

[13] Huberty, Carl J., E. Michael Jordan, and W. Christopher Brandt. "Cluster analysis in higher education research." In Higher education: Handbook of theory and research, pp. 437-457. Springer, Dordrecht, 2005.

[14] Moucary, C. El, Marie Khair, and Walid Zakhem. "Improving student's performance using data clustering and neural networks in foreign-language based higher education." The Research Bulletin of Jordan ACM 2, no. 3 (2011): 27-34.

[15] Singh, Ishwank, A. Sai Sabitha, and Abhay Bansal. "Student performance analysis using clustering algorithm." In 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence), pp. 294-299. IEEE, 2016.

[16] De Morais, Alana M., Joseana MFR Araujo, and Evandro B. Costa. "Monitoring student performance using data clustering and predictive modelling." In 2014 IEEE Frontiers in Education Conference (FIE) Proceedings, pp. 1-8. IEEE, 2014.