International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

An Unsupervised Ensemble Clustering Approach for the Analysis of Student Behavioral Patterns

G. Varna[1], L. Poorvi[2], V. Sai Priya[3]

Department of Computer Science and Engineering, Stanley College of Engineering and Technology for Women, Telangana, India

Abstract.
Specialized services and management must understand students' behavioural patterns in a timely and accurate manner. Based on these patterns, we can make targeted rules, especially for unexpected patterns. To perform this type of work, a questionnaire method is usually used to collect data and analyse students' behavioural states. However, the effectiveness of this method is greatly influenced by the timeliness and validity of the feedback data. To address this problem, we propose an unsupervised ensemble clustering framework to use student behavioural data to discover behavioural patterns. Because the behavioural data produced by students on campus are available in real time without intentional bias, clustering analysis can be relatively efficient and reliable. The proposed framework extracts behaviour features from the two perspectives of statistics and entropy and then combines density-based spatial clustering of applications with noise (DBSCAN) and k-means algorithms to discover behavioural patterns. To evaluate the performance of the proposed framework, we carry out experiments on six types of behavioural data produced by undergraduates in a university in Beijing and analyse the relationships between different behavioural patterns and students' grade point averages (GPAs). The results show that the framework can not only detect anomalous behavioural patterns but also find mainstream patterns. The findings from this research can assist student-related departments in providing better services and management, such as psychological consulting and academic guidance.

1. Introduction

1.1 About Project

An important task in the education field is discovering student behavioural patterns and taking the corresponding actions to optimize the educational process—for example, finding various behavioural factors that have strong correlations with academic performance, analyzing student learning behaviours to allow teachers to adjust teaching schedules for better outcomes and to give early warnings to students who may fail exams, modelling the mobility flow of students on campus to support the reasonable allocation of resources by administrators, detecting students' anomalous behaviours so that managers can take timely preventive measures, and determining social networks from behavioural data to identify solitary students.

1.2 Objectives of the Project

Supervised approaches require labelled student data and the training of a classification model to determine which class an unseen student belongs to. Semi supervised approaches build a model to learn the representative features of students who belong to only one class. A student is marked as not belonging to the class when the

difference between his or her features and the representative features exceeds the specified threshold. However, labelled student data, especially that of anomalous students, are not available because of privacy concerns. Additionally, student labels keep evolving, which means any model must be updated dynamically. These factors make supervised and semi supervised approaches difficult to apply in practice. In contrast, unsupervised approaches do not require labels and fully exploit the nature of datasets to cluster instances, so they are widely used in practical applications.

1.3 Scope of the Project

Our main contributions are summarized as follows. 1) Six types of behavioral data in time series format were collected, and the features for each type of behavior are extracted from the perspectives of central tendency, dispersion and entropy, which provides a more reliable basis for the analysis of behavioral patterns. 2) An ensemble unsupervised clustering framework is proposed by fully taking advantage of the DBSCAN and k-means algorithms; this framework can detect unexpected behavioral patterns and discover mainstream behavioral patterns. The clustering results provide helpful information for specialized management. 3) GPA levels are taken as the ground truth to calculate extrinsic metrics to measure the correlation between different behaviors and academic performance.

2. Literature Survey

2.1 Existing System

Most researchers use a questionnaire survey method to collect data from specific students in specific circumstances. However, the method of collecting data has some limitations. First, it is impossible to capture students' current state in a timely manner withthis method because surveys are conducted on a scheduled basis, such as one per academic year or semester. If students with unexpected behavioral patterns cannot be identified in a timely manner, there may be serious consequences. Second, students exhibiting anomalous behaviors may deliberately supply false information to make them appear normal, while normal students may not carefully fill out the survey,so the collected data could contain noise or false informationthat bias the analysis results. Third, rich expert knowledge is needed to design a questionnaire that can capture enough information to comprehensively analyze students' behavioral patterns.

2.2 Proposed System

We propose an unsupervised ensemble clustering framework to use student behavioral data to discover behavioral patterns. Because the behavioral data produced by students on campus are available in real time without intentional bias, clustering analysis can be relatively efficient and reliable. The proposed framework extracts behavior features from the two perspectives of statistics and entropy and then combines density-based spatial clustering of applications with noise (DBSCAN) and k-means algorithms to discover behavioral patterns.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

To evaluate the performance of the proposed framework, we carry out experiments on six types of behavioral data produced by undergraduates in a university in Beijing and analyze the relationships between different behavioral patterns and students' grade point averages (GPAs)

3. System Architecture

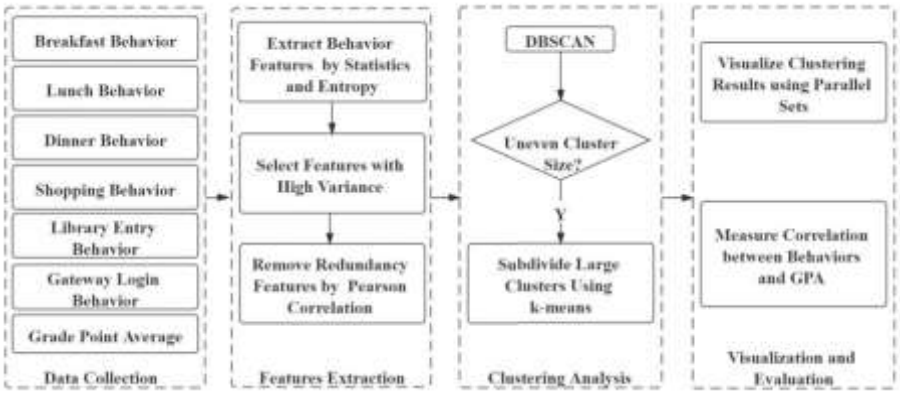

Fig.1. System Architecture

4. Implementation

4.1 Algorithm

A.CLUSTERING RESULTS USING DBSCAN

To determine the parameters Eps and MinPts of DBSCAN, we plot a MinPts-dist graph for each type of behavior, where MinPts is set from 2 to 24. In the six graphs, the curves do not significantly change when MinPts is greater than 8, so we set MinPts to 8. The 8-dist graphs show that the optimal Eps values are 0.231 for breakfast behavior, 0.14 for lunch behavior, 0.175 for dinner behavior, 0.124 for shopping behavior, 0.082 for library entry behavior, and 0.09 for gateway login behavior. The clustering results of DBSCAN with the given values of Eps and MinPts are shown in Figs. 6 and 7, where -1 is the label of the noise cluster, the normal clusters are labeled with numbers starting from 0, and the number of students in each cluster is above its bar. For example, there are a total of 19 clusters numbered from -1 to 17 for breakfast behavior, as shown in Fig. 6(a); noise cluster -1 contains 184 students who can be identified as those with unexpected behavioral patterns; clusters numbered 2, 4, 12, 13, 14, 15, 16 and 17 all contain relatively few students, less than 200, so the behavioral patterns they represent should be in the minority; clusters 0, 1, 3, 5, 6, 7, 8, 9, 10 and 11 all contain relatively large numbers of students, and they can represent students' mainstream behavioral patterns, especially clusters 0, 1, and 3. Based on the results, student services and management departments should pay more attention to the noise clusters and minority clusters for early warnings and provide targeted services and management according to mainstream patterns.

B. KMeans:

The final clustering results of these four types of behaviors after subdividing cluster 0 with the given k are shown in Figs. 9 and 10, in which the clusters suffixed with 'DBSCAN' are noise clusters and minority clusters generated by DBSCAN, while clusters suffixed with 'KMEANS' are the subclusters subdivided using k-means. The number of students in each cluster is above the bar. The final result not only retains the noise and small clusters but also subdivides the large clusters into basically uniform sub clusters.

4.2 Code Implementation

Python 3.7. Python is broadly utilized universally and is a high-level programming language. It was primarily introduced for prominence on code, and its language structure enables software engineers to express ideas in fewer lines of code. Python is a programming language that gives you a chance to work rapidly and coordinate frameworks more effectively. The working of the code involves several steps, they are as follows:

- To run project double click on 'run.bat' file
- click on 'Data Collection/Upload Dataset' button to upload dataset
- select and upload dataset file
- dataset loaded but it contains lots of non-numeric values so click on 'Features Extraction' button to convert non-numeric values to numeric and then apply PCA to select important features
- we can see all values are converted to numeric and we can see dataset contains total 17 columns or attributes and after applying PCA we got 6 features and the selected 6 features names you can see on the screen and while applying PCA we got feature variance graph for each features
- In the graph x-axis represents features names and y-axis represents importance value between 0 and 1 and if feature is important then its value will be closer to 1 else 0. Now close above graph and click on 'Run Initial DBSCAN Clustering' button to perform clustering
- Now the graph represents small black dots are the noise cluster records and big dots are main single cluster which contains normal behaviour and now close above graph
- We can see total records contains in dataset and total number of records group into DBSCAN main cluster and total records into noise clusters and in above screen we can see record ID'S of abnormal behaviour students and now click on 'Run KMEANS Clustering' button to apply KMEANS to main single DBSCAN cluster to get below graph
- In KMEANS graph we can see 10 different colours dots which means 10 different clusters are created and then cluster with least students will be consider as ANOMALOUS and now click on 'Visualize Clusters' graph to get graph of KMEANS

- In KMEANS graph x-axis represents cluster no and y-axis represents number of students in that cluster and the cluster with least students will be identified as ANOMALOUS and now close above graph and click on 'Display Anomalous Student ID's' button to get all anomalous students ID
- In result we can see 3 different clusters in square brackets which contains least records and consider as ANOMLAOUS so out of 10 clusters above 3 clusters are ANOMALOUS or noise cluster. Inside square bracket we can see student id and this id you can see dataset record numbers

5.Result

Fig 5.a Output screen

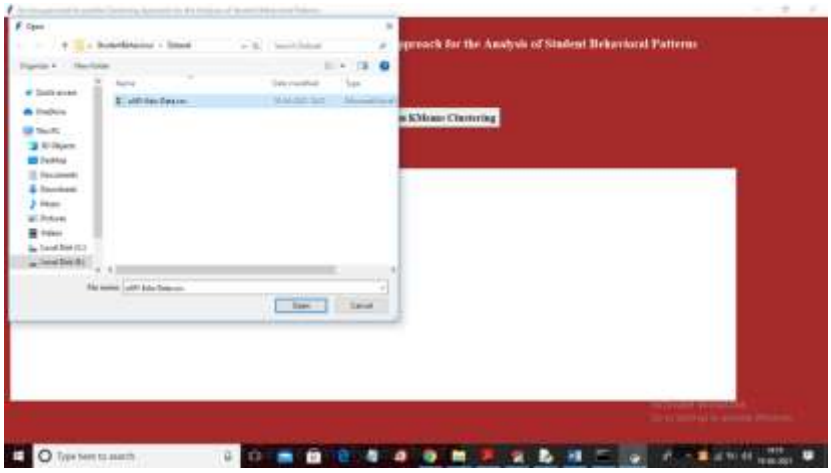

Fig 5.b Uploading data set

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

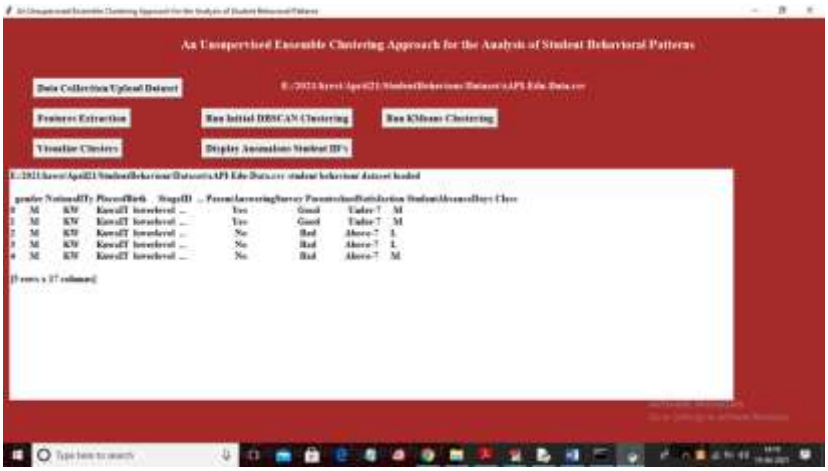

Fig 5.c Dataset uploaded

Fig 5.2 Feature Extraction

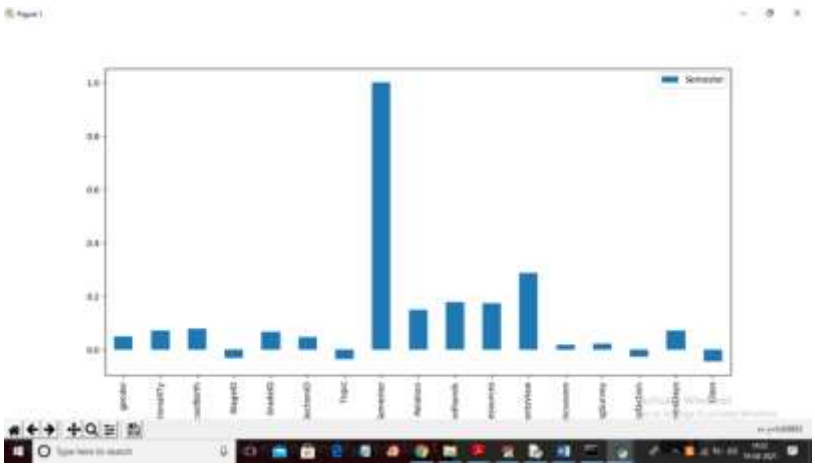

Fig 5.d Feature Variance graph

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

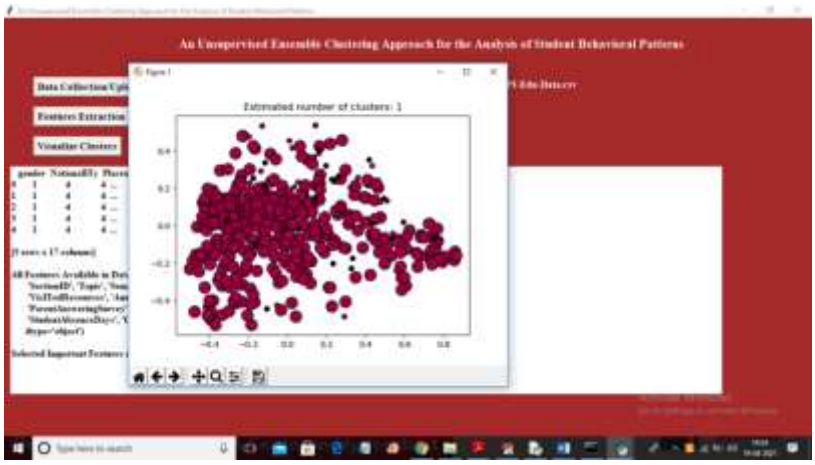

Fig 5.2 Initial DBSCAN clustering result

Fig 5.e Anomalous student id result after DBSCAN

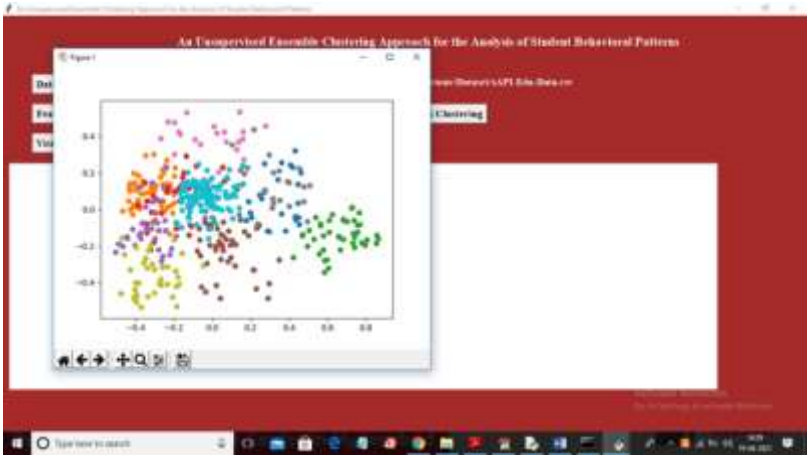

Fig 5.f KMEANS graph

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

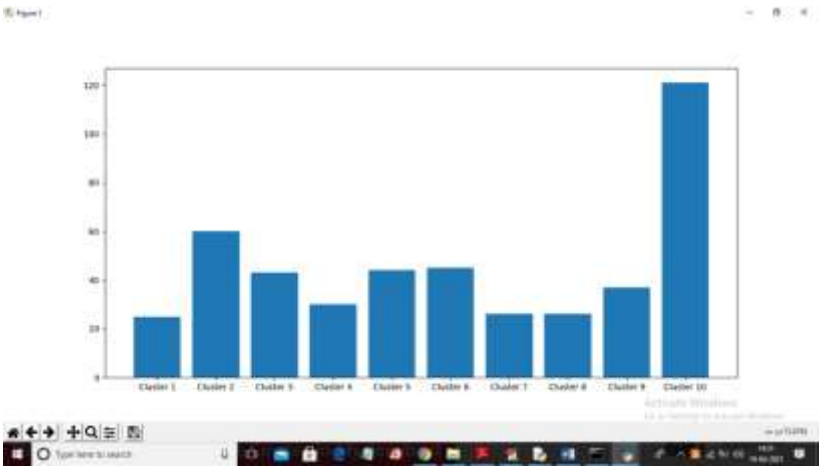

Fig 5.2 graph of KMEANS

Fig 5.2 Result

7.Conclusion

This paper proposed an ensemble unsupervised clustering framework for the analysis of students' behavioural patterns by combining DBSCAN and k-means algorithms. To evaluate the effect of the proposed method, we collect six types of behavioural data produced by 9024 undergraduates on campus and extract behavioural features through the two aspects of statistics and entropy. The experimental results demonstrate that the proposed method can not only detect anomalous behavioural patterns but also more precisely identify mainstream behavioural patterns. Based on the clustering results, student departments can adopt more targeted

measures for intervention and specialized services. At the end of the paper, we discuss three issues: whether we can cluster the behavioural feature space using only the k-means algorithm, the difficulty of applying the proposed method to high-dimensional multisource behaviour features, and the relationship between different behavioural patterns and academic performance levels.

8. Future Scope

For better clustering analysis, future work should include the following:
(1) Extract more meaningful features by fully fusing multisource behavioural data
(2) Design a new distance measure to make the proposed method effective for high dimensional feature spaces
(3) Further study the relationship between behavioural patterns and student labels, such as academic performance, psychological state, and employment domain.

9. References

1. [1] Programming Python, Mark Lutz
2. [2] Head First Python, Paul Barry
3. [3] Core Python Programming, R. Nageswara Rao
4. [4] Learning with Python, Allen B. Downey
5. [5] A. H. Eliasson, C. J. Lettieri, and A. H. Eliasson, ‘‘Early to bed, early to rise! Sleep habits and academic performance in college students,’’ Sleep Breathing, vol. 14, no. 1, pp. 71–75, Feb. 2010, doi: 10.1007/s11325-009- 0282-2.
6. [6] X. D. Keating, D. Castelli, and S. F. Ayers, ‘‘Association of weekly strength exercise frequency and academic performance among students at a large university in the united states,’’ J. Strength Conditioning Res., vol. 27, no. 7, pp. 1988–1993, Jul. 2013, doi: 10.1519/JSC.0b013e318276bb4c.
7. [7] M. Valladares, E. Duran, A. Matheus, S. Duran-Agueero, A. M. Obregon, and R. Ramirez-Tagle, ‘‘Association between eating behavior and academic performance in university students,’’ J. Amer. College Nutrition, vol. 35, no. 8, pp. 699–703, 2016, doi: 10.1080/07315724.2016.1157526.
8. [8] https://www.w3schools.com/python/
9. [9] https://www.tutorialspoint.com/python/index.htm

10. [10] https://www.javatpoint.com/python-tutorial

11. [11] https://www.learnpython.org/

12. [12]https://www.pythontutorial.net/
13. Kishor Kumar Reddy C and Vijaya Babu B, "ISPM: Improved Snow Prediction Model to Nowcast the Presence of Snow/No-Snow", International Review on Computers and Software, 2015.
14. (http://www.praiseworthyprize.org/jsm/index.php?journal=irecos&page=article&op=view&path%5B%5D=17055)
15. Kishor Kumar Reddy C, Rupa C H and Vijaya Babu B, "SLGAS: Supervised Learning using Gain Ratio as Attribute Selection Measure to Nowcast Snow/No-Snow", International Review on Computers and Software, 2015.
16. (http://www.praiseworthyprize.org/jsm/index.php?journal=irecos&page=article&op=view&path%5B%5D=16706)
17. Kishor Kumar Reddy C, Vijaya Babu B, Rupa C H, "SLEAS: Supervised Learning using Entropy as Attribute Selection Measure", International Journal of Engineering and Technology, 2014.
18. (http://www.enggjournals.com/ijet/docs/IJET14-06-05-210.pdf)
19. Kishor Kumar Reddy C, Rupa C H and Vijaya Babu B, "A Pragmatic Methodology to Predict the Presence of Snow/No-Snow using Supervised Learning Methodologies", International Journal of Applied Engineering Research, 2014.
20. (http://www.ripublication.com/Volume/ijaerv9n21.htm)
21. Kishor Kumar Reddy C, Rupa C H and Vijaya Babu, "SPM: A Fast and Scalable Model for Predicting Snow/No-Snow", World Applied Sciences Journal, 2014.
22. (http://www.idosi.org/wasj/wasj32(8)14/14.pdf)
23. Kishor Kumar Reddy C, Anisha P R, Narasimha Prasad L V and Dr. B Vijaya Babu, "Comparison of HAAR, DB, SYM and COIF Wavelet Transforms in the Detection of Earthquakes Using Seismic Signals", International Journal of Applied Engineering Research, 2014, pp. 5439-5452.