# Innovative Teaching-Learning Process: Categorical Clustering Data

**K.SreeVani**
Assistant Professor
Vidya Jyothi Institute of Technology,Hyderabad, Telangana State.
sreevani@vjit.ac.in

**Abstract:** Clustering is process, grouping a set of physical or abstract objects into classes of similar objects. Clustering techniques can be broadly classified into many categories; partitioning, hierarchical, density-based, grid-based, model-based algorithms. The present study is intended to explore the *categorical clustering data.* The objectives of the study were to explore the *levels* of categorical data clustering among the students pursuing Engineering courses in Hyderabad District of Telangana State with special reference to gender. A self-developed questionnaire was administered on the selected sample of *one hundred and eighty students* pursuing Engineering courses. The results revealed that there is a statistically significant difference in categorical data clustering with reference to gender as well as managementImplications and suggestions for further research were also portrayed.
**Key words**: Clustering; categorical data; clustering techniques; partitioning; hierarchical; density-based; grid-based; and model-based algorithms.

## INNOVATIVE TEACHING-LEARNING PROCESS: CATEGORICAL CLUSTERING DATA

### 1.0 Introduction

Clustering is a process, grouping a set of physical or abstract objects into classes of similar objects. Clustering is an unsupervised machine learning techniqueused to group unlabelleddata into clusters that contain datapoints that are 'similar' to each other and 'dissimilar' fromthose in other clusters (Jain,1988 & Khan,2004).It is called as unsupervised learning because it does not use predefined classes or labels for clustering data (Sowmiya, 2017).

The classical definition of cluster was attributed by M. Porter (2008): "Educational cluster is a group of geographically neighbouring interconnected companies and organizations connected to them, working in a certain area and characterized by common activities and mutual reinforcement".

"Educational cluster" is a complex of educational institutions of all educational levels, industries of correspondent government bodies, whose activity is connected with industries and is aimed at successful innovative development" (Galimova, 2009). "Educational cluster" is a complex of interconnected institutions of vocational education, connected branch-wise and by partnership with the industry players (Zhuravlyova&Bashkirtseva, 2008).

Clustering of categorical data is becoming increasingly important, since non-numerical data are ubiquitous and clustering can be used, for example, to optimize an anonymization process or to perform anomaly detection, or in any application where there is the need to automatically recognize the intrinsic structure of data.

The data containing categorical attributes pose a number of challenges on the existing clustering methods due to a) no natural order; b) high dimensionality; c) existence of sub-space clusters and d) conversion of categorical to numerical data.

Clustering techniques can be broadly classified into many categories; partitioning, hierarchical, density-based, grid-based, model-based algorithms.

The educational system in India is currently facing several issues such as identifying students need, personalization of training and predicting quality of student interactions. Educational data mining (EDM) provides a set of techniques which can help educational system to overcome this issue in order to improve Learning experience of students as well as increase their profits (Veeramuthu, 2014).

Categorical clustering is a new phenomenon in the field of education and a few studies were conducted in India viz.,Khandelwal and Sharma (2015) proposed a fastcategorical clustering algorithm; Sharma and Gaud (2015) modifications in theclassic K-modes algorithm; Ahmad and Khan (2013) addressed this initialization problem of K-modes algorithm; Goswami and Mohanta (2004) have proposed aclustering approach using the distance metric; Ibrahim and Harbi(2012) proposed a Modified PAM (MPAM) clustering algorithm. In other words, research in this field is in an embryonic stage. Moreover, studies conducted at local level seem to be a distance dream. Hence, the present study is stated as:

## INNOVATIVE TEACHING-LEARNING PROCESS: CATEGORICAL CLUSTERING DATA

### 1.1 Objectives of the study

1. To explore the *levels* of categorical data clustering among the Engineering Students in Hyderabad District of Telangana State.
2. To analyze the categorical data clustering among the Engineering Students in Hyderabad District of Telangana State with reference to the *gender.*
3. To study the categorical data clustering among the Engineering Students in Hyderabad District of Telangana State with special reference to *management.*

### 1.2 Delimitations of the Study

The study is confined to investigate the categorical data clustering with regard to gender and management of the

Engineering Students in Hyderabad District of Telangana State.

### 1.3 Previous Literature

In a nutshell, He, Z., Xu, X., Deng, S., & Huang, J. Z. (2004) proposed an efficient clustering algorithm for analyzing categorical data streams; Kotsiantis et al. (2004) propounded five classification algorithms; Romero, C., & Ventura, S. (2007) unearthed application of data mining to traditional educational systems; IndrajitSaha and Anirban Mukhopadhyay (2008) demonstrated a variety of artificial and real life categorical data sets.; Do, H. J., & Kim, J. Y. (2008) indicated a new clustering algorithm for categorical data; Yu et al (2010) explored student retention by using classification trees; Ramaswami and Bhaskaran (2010) focused on developing predictive data mining model to identify the slow learners; Aranganayagi, S., & Thangavel, K. (2010) presented an incremental algorithm to cluster the categorical data; Sayal, R., & Kumar, V. V. (2011) overviewed of popular similarity measures of categorical attributes; Rezankova, H., Loster, T., &Husek, D. (2011) studied criteria based on variability measures; Baradwaj, B. K., & Pal, S. (2012) attempted data mining techniques in context of higher education; Kalaivani, K., &Raghavendra, A. P. V. (2012) presented categorical data set; Md.Hedayetul Islam Shovon (2012) presented a paper on prediction of student academic performance by applying K-means clustering; Sisodia, D., Singh, L., Sisodia, S., &Saxena, K. (2012) dealt with the study of various clustering algorithms of data mining; SwastiSinghal, Monika Jena (2013) introduced the WEKA tool; Chandrika, J., & Kumar, K. A. (2013) delineated cluster the transactional data streams; Venkatesan, N. (2013) discussed the types of modeling technique; Kabakchieva, (2013) high potential of data mining applications for university management; Durairaj et al., (2014) proposed Educational Data mining; Natek, Srečko, and MotiZwilling., (2014) focused on the study of data mining techniques; Prashant et al. (2014) examined the clustering analysis in data mining; Veeramuthu et al (2014) analyzed how different factor affect a Students learning behavior; Shiwani and Roopali (2016) applied unsupervised learning algorithms; Gul'zamira D. Aitbayeva et al (2016) studied educational clusters; Sowmiya and Valarmathi. (2017) presented the literature review of the clustering algorithm for categorical and binary attributes; Abdul Rahmat (2017) studied transformational intellectual; Uddin J, Ghazali R, Deris MM (2017) proposed an alternative technique named Maximum Indiscernible Attribute (MIA); Sangam, R. S., & Om, H. (2017) proposed k-mode stream; Govindasamy, K., &Velmurugan, T. (2018) studied four clustering algorithms; Lakshmi Sreenivasa Reddy and Rajini (2018) strategic management tool; Qin, H., & Ma, X. (2018) propounded IG-ANMI; Amir Ahmad and SherozS.Khan (2019) presented taxonomy for the study of mixed data clustering algorithms.

### 1.4 Operational Definitions

Operational definitions define concepts in terms of operations or process.

**a) Categorical data**

In the present study, categorical data refers to40 words divided into 4 categories (flowers, fruits, animals and cities) with 10 words in each category.

**b) Clustering**

In the present context, clustering denotes grouping of words into each category according to a similar property.

### 1.5 Sample

In order to select the representative sample for the study, simple random sampling technique was used. **One hundred and eighty students**, (boys and girls) from the Engineering Students from University as well as from autonomous (private) colleges in Hyderabad District of Telangana Statewere selected for the present investigation.

### 1.6 Instrumentation

Table No.1 shows individual data for the total number of categories formed and the range of words in those categories

| **Recall Tests** | **No. of Categories formed and No. of words recalled** | **No. of clusters recalled and (No. of words recalled under each cluster)** | | | | **Total clusters** |
|---|---|---|---|---|---|---|
| | | I | II | III | IV | |
| | | vocabulary | grammar | listening | speaking | |
| Recall Test -1 | 3(21) | 0 (0) | 1(6) | 1(5) | 2(8) | 4 |
| Recall Test -2 | 4(26) | 1(3) | 1(3) | 4(9) | 2(8) | 8 |
| Recall Test -3 | 4(33) | 2(9) | 2(4) | 2(9) | 1(8) | 7 |

### 1.7 Results and Discussion

**$HO_1$**. There is no statistically significant difference on the levels of categorical data clustering among Engineeering Students in Hyderabad District of Telangana State.

Table 4.1 showing mean scores and ANOVA on the levels of Categorical data clustering

| | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Minimum | Maximum |
|---|---|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound | | |
| Boys (Uni) | 45 | 24.4444 | 7.77688 | 1.15931 | 22.1080 | 26.7809 | 8.00 | 38.00 |
| Girls (Uni) | 45 | 20.1333 | 7.47298 | 1.11401 | 17.8882 | 22.3785 | 8.00 | 39.00 |
| Boys (Pvt) | 45 | 25.3556 | 6.97123 | 1.03921 | 23.2612 | 27.4499 | 11.00 | 38.00 |
| Girls (Pvt) | 45 | 20.5778 | 7.32830 | 1.09244 | 18.3761 | 22.7794 | 8.00 | 39.00 |
| **Total** | **180** | **22.6278** | **7.68517** | **.57282** | **21.4974** | **23.7581** | **8.00** | **39.00** |

It can be observed from the ANOVA table the calculated p-value is 0.001, which is highly significant. Moreover, the mean score of Private ITI Boys have is 25.3556 ± 6.97123 followed by Government ITI boys 24.4444 ±7.77688. Thus, it can be inferred that there is a statistically significant difference on the levels of categorical data clustering among Engineering students in Hyderabad District of Telangana State. Hence, the hypothesis formulated was **rejected**.

**HO2**: There is no statistically significant difference on the categorical data clustering among Engineering students of Hyderabad District of Telangana State with reference to *gender*.

Table 4.2 showing mean scores and t-test on the Categorical Clustering.

| Gender | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|
| Boys (Private) | 45 | 25.3556 | 6.97123 | 1.03921 |
| Girls (Private) | 45 | 20.5778 | 7.32830 | 1.09244 |
| Boys (Univ) | 45 | 24.4444 | 7.77688 | 1.15931 |
| Girls (Univ) | 45 | 20.1333 | 7.47298 | 1.11401 |

The t-test table demonstrates the calculated mean ± Sd is found to be very high in boys when compared to girls. The descriptive statistics shows 24.4444 ± 7.77688 (University Engineering Boys) and 25.3556 ± 6.97123 (Private, engineering students). Thus, the hypothesis was **rejected.**

**HO3**: There is no statistically significant difference on the categorical data clustering among Engineering students of Hyderabad District of Telangana State with reference to *management.*

Table 4.3 showing mean scores and t-test on the Categorical data clustering

| Gender | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|
| Boys (Private) | 45 | 25.3556 | 6.97123 | 1.03921 |
| Girls (Private) | 45 | 20.5778 | 7.32830 | 1.09244 |
| Boys (Univ) | 45 | 24.4444 | 7.77688 | 1.15931 |
| Girls (Univ) | 45 | 20.1333 | 7.47298 | 1.11401 |

The Mean ± Sd is found to be very high in boys when compared to girls. The descriptive statistics shows it is 24.4444 ± 7.77688 (University Engineering Boys) and 25.3556 ± 6.97123 (Private Engineering Boys). Hence, it can be deduced that there is a statistically significant difference on the levels of categorical data clustering among Engineering students in Hyderabad District of Telangana State with reference to management. Thus, the hypothesis was **rejected.**

### 1.7 Major Findings

1. The present study revealed that there is a statistically significant difference on the levels of categorical data clustering among theEngineering Students in Hyderabad District of Telangana State.
2. The findings of the study explicitly demonstrated there is a statistically significant difference on the levels of categorical data clustering among the Engineering Students in Hyderabad District of Telangana Statewith reference to gender.
3.The results illuminated a statistically significant difference on the levels of categorical data clustering among Engineering students in Hyderabad District of Telangana State with reference to management.

### 5.4 Implications of the Study

The current research exhibited sanguine implications for teachers, young researchers, administrators and also for Policy Makers. The present piece of research proposes to ***teachers*** that an alternative representation of categorical data as numeric data making it easier to handle. This technique provides a uniform representation for data points and the cluster representatives. In the same manner, ***young researchers*** may measure psychological, aptitude, and achievement characteristics. A cluster analysis then may identify what homogeneous groups exist among students (for example, high achievers in all subjects, or students that excel in certain subjects but fail in others).As clustering seeks patterns in educational datasets, it holds implications for ***administrators/authorities*** to come out with broad-based curricular adaptations in the field of education. The ***Policy makers*** may focus on categorical data variables characterized by values, which are classified into: dichotomous, multi-categorical.

### 5.5 Suggestions for Further Research

The present study is not much comprehensive and exhaustive due to its limitations. Thus, it is suggested

that further investigations may be focused on the following issues:

1. A similar study can be conducted with a *larger group of respondents* to have in- depth knowledge on the clustering categorical data.
2. There is a need to explore the *categorical clustering data* in achieving quality education.
3. An explorative study can be taken up on *students' academic performance* through mining educational data
4. A study can be taken on the *factors that influence efficacy of clustering categorical data.*
5. Another area for investigation would be on *analysis of student result using clustering techniques.*

**BIBLIOGRAPHY**


1. Aranganayagi, S., &Thangavel, K. (2010). Incremental algorithm to cluster the categorical data with frequency based similarity measure.*World Academy of Science, Engineering and Technology*. Vol:4 2010-01-23
2. Baradwaj, B. K., & Pal, S. (2012). Mining Educational Data to analyze students' performance. *arXiv preprint arXiv:1201.3417*.
3. Chandrika, J., & Kumar, K. A. (2013).A Novel Approach for Clustering Categorical Data Streams. *International Journal of Innovation, Management and Technology*, *4*(5), 486.
4. Do, H. J., & Kim, J. Y. (2008), FAVC: Clustering Categorical Data Using the Frequency of Attribute Values Combinations. In *2008 3rd International Conference on Innovative Computing Information and Control* (pp. 304-304).IEEE.
5. Govindasamy, K., &Velmurugan, T. (2018).Analysis of Student Academic Performance Using Clustering Techniques. *International Journal of Pure and Applied Mathematics*, *119*(15), 309-323.
6. He, Z., Xu, X., Deng, S., & Huang, J. Z. (2004). Clustering Categorical data streams. *arXiv preprint cs/0412058*.
7. IndrajitSaha andAnirbanMukhopadhyay (2008) Improved Crisp and Fuzzy Clustering Techniques for Categorical Data. *International Journal of computer Sciences*. 35 (4).On line publication.
8. Kalaivani, K., &Raghavendra, A. P. V. (2012). Efficiency based categorical data clustering. In *2012 IEEE International Conference on Computational Intelligence and Computing Research* (pp. 1-4). IEEE.
9. Kotsiantis, S., C. Pierrakeas, P. Pintelas. Prediction of Student's Performance in Distance Learning Using Machine Learning Techniques. Applied Artificial Intelligence,Vol.18,2004,No5,411-426.
10. Lakshmi Sreenivasa Reddy and Rajini (2018)A Proposal to Predict Student's Performance using Data Mining Techniques.*International Journal of Computer & Mathematical Sciences*.IJCMS.Volume 7, Issue 2.pp.97-101.
11. Md. Hedayetul Islam Shovon,(2012) Prediction of Student Academic Performance by an Application of K-Means Clustering Algorithm, *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 2(7).
12. Qin, H., & Ma, X. (2018). IG-ANMI: a novel initialization method for genetic clustering algorithm for categorical data. *International Journal of Science, Engineering and Technology (IJSET) UTY*, *1*(1), 53-66.
13. Rahamat (2017) Clustering in Education.*European Research Studies Journal* Volume XX, Issue 3A, 2017 pp. 311-324.
14. Ramaswami, M., R. Bhaskaran. A CHAID Based Performance Prediction Model in Educational Data Mining. – IJCSI *International Journal of Computer Science* Issues, Vol. 7, January 2010, Issue 1, No 1, 10-18.
15. Sangam, R. S., & Om, H. (2017). K-modestream algorithm for clustering categorical data streams. *CSI Transactions on ICT*, *5*(3), 295-303.
16. Sayal, R., & Kumar, V. V. (2011).A novel similarity measure for clustering categorical data sets. *International Journal of Computer Applications*, *17*(1), 25-30.
17. Shruti Sharma, Manoj Singh (2015) Clustering with Categorical Data-A Survey.International Journal of Engineering, Management & Sciences (IJEMS). Volume-2, Issue-12,pp.1-5.
18. Sisodia, D., Singh, L., Sisodia, S., &Saxena, K. (2012). Clustering techniques: a brief survey of different clustering algorithms. *International Journal of Latest Trends in Engineering and Technology (IJLTET)*, *1*(3), 82-87.
19. Sowmiya, N., &Valarmathi, B.(2007) A Review of categorical data clustering methodologies based on recent studies.*IIOABJ*. Vol. 8 (2.) pp.353-365.
20. Uddin J, Ghazali R, Deris MM (2017) An Empirical Analysis of Rough Set Categorical Clustering Techniques. *PLoS ONE* 12(1): e0164803.
21. Veeramuthu, P., Periyasamy, D. R., &Sugasini, V. (2014).Analysis of student result using clustering techniques. *International Journal of Computer Science and Information Technologies*, *5*(4), 5092-5094.
22. Venkatesan, N. (2013). Role of Data Mining Techniques in Educational and E-learning System. *Asia Pacific Journal of Research*, *2*.
23. Yu, C., S. DiGangi, A. JannaschPennell, C. Kaprolet.(2010) A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year. *Journal of Data Science*, Vol. 8,pp. 307-325.


****