



## CLUSTERING ALGORITHMS IN EDUCATIONAL DATA MINING: A REVIEW

C.Anuradha<sup>1</sup>, T.Velmurugan<sup>2</sup>, R. Anandavally<sup>3</sup>

<sup>1</sup>Research Scholar, Bharathiar University, Coimbatore, India.

<sup>2</sup>Associate Professor, PG and Research Dept. of Computer Science, D.G.Vaishnav College, Chennai-106.

<sup>3</sup>Asst.Professor, Dept. of Computer Science, SSBSTAS College, Tamilnadu.

<sup>1</sup>[anumphil14@gmail.com](mailto:anumphil14@gmail.com); <sup>2</sup>[velmurugan\\_dgvc@yahoo.co.in](mailto:velmurugan_dgvc@yahoo.co.in); <sup>3</sup>[anandhi05@gmail.com](mailto:anandhi05@gmail.com)

### Abstract

Currently, universities record large amounts of data about students. Despite its potential to inform decisions regarding the allocation of resources and efforts, this information tends to be overlooked. Educational data mining is a recent research field that focuses on the use of data mining techniques to transform large volumes of educational data into useful and relevant knowledge that can improve the educational processes and decisions. Among the many data mining techniques, clustering helps to classify the student in a well-defined cluster to find the behavior and learning style of students. The primary objective of this review paper will portray how the clustering techniques classify the student in a well-defined cluster which enables academicians to predict student's final results according to their academic performance at a first stage. Furthermore, this research paper focuses on application of clustering in EDM done by various researchers. Additionally, it endows academicians to estimating their students capabilities and follow up by forming a new system entry.

**Keywords:** Data mining, Educational data mining, Clustering, K-means, Hierarchical, Student performance

### 1. INTRODUCTION

Education is essential for country's development. Education has the ability to change and to induce change and progress in society. In the last decade it has been conducted a deep analysis, particularly on higher Education, which forced the evaluation, review and reformulation of the processes used to guarantee the quality of the education services provided. Data mining is a computer-assisted

process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining techniques predict behaviors and future trends, allowing decision makers to make proactive, knowledge-driven decisions. Some of the most useful techniques include statistical methods, data visualization, association rule mining, classification and clustering [1]. Having a better understanding of which students are more likely to

face difficulties in their educational process and identifying the factors that influence these difficulties, higher education institutions will be able to timely develop strategies to increase the graduation rate and mitigate their attrition rates.

Clustering analysis is one of the techniques used in data mining and involves the partitioning of a set of data objects into subsets. These subsets or clusters are used to organize objects in such a way that each object within in a cluster is similar to one another yet they are dissimilar to other clusters [2]. Among the various clustering techniques, number of studies have utilized K-means algorithm on the basis of the ease of use, simplicity and performance of the algorithm [3]. This research work aims at inspecting the use of educational data in clustering techniques such as social-economic, demographic, higher education access average and academic results to identify bottlenecks that control academic success and to predict students academic performance.

This research paper focuses on set up a clustering algorithm which is most suitable for predicting student performance in Educational data mining. The objective of this research work is to gain an insight into how clustering analysis can be done in educational domain and to highlight the potential characteristics of the clustering algorithms within the educational data set.

This paper is organized as follows. Section 2 sketches a general review of the families of clustering techniques along with a description is given. Section 3 describes the clustering algorithms widely used in educational data mining. The literature survey of the research on clustering techniques in education is given in Section 3. Finally Section 4 concludes by summarizing their findings.

## 2. CLUSTERING TECHNIQUES

Clustering algorithms can be classified by the data type analyzed, similarity measure used to group data and the theory used to define the cluster [4]. Some of the clustering methods are commonly used and are available in various data mining tools such WEKA. These techniques and data mining tools are also commonly used in educational data mining analysis [5]. This section highlights the various types of clustering techniques available in data mining.

### 2.1 Clustering Methods

**Partitioning Method:** The partitioning methods generally result in a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster. The precise form of this description will depend on the type of the object which is being clustered. In case where real-valued data is available, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases, e.g., a cluster of documents can be represented by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of the clusters is large, the centroid can be further clustered to produces hierarchy within a dataset [6].

**Single Pass:** A very simple partition method, the single pass method creates a partitioned dataset as follows.

1. Make the first object the centroid for the first cluster.
2. For the next object, calculate the similarity,  $S$ , with each existing cluster centroid, using some similarity coefficient.

3. If the highest calculated S is greater than some specified threshold value, add the object to the corresponding cluster and re determine the centroid; otherwise, use the object to initiate a new cluster. If any objects remain to be clustered, return to step 2.

**Hierarchical methods:** Hierarchical-clustering methods group data objects into a hierarchy or tree of clusters. Hierarchical clustering initializes a cluster system as a set of singleton clusters (agglomerative case) or a single cluster of all points (divisive case) and proceeds iteratively merging or splitting the most appropriate clusters until the stopping criterion is achieved [7].

**Density based methods:** Density-based methods regard clusters as dense regions in the data space separated by sparser regions. Some of these methods approximate the overall density function, like mixture model methods, while others use only local density information to construct clusters. In this sense, the mixture model methods could also be classified as (parametric) density-based methods, although the term usually refers to non-parametric methods. The advantage of the non-parametric density-based methods is that one does not have known the number or distributional form of the sub clusters [8].

### 3. CLUSTERING ALGORITHMS

One of the pre-processing algorithms of EDM is known as Clustering. This Section discuss about the clustering algorithms which can be widely used in educational data mining. The methods described here can be applied by various researchers to predict academic performance of students.

#### 3.1 K-Means Clustering Algorithm

Big data sets can be easily clustered by using K means clustering algorithm. The main advantage of

clustering data set is to reduce the time required to process and access each dimension. K data elements are selected as initial centers and Euclidean distance formula is used to calculate distance between the selected centroid and other data elements and then the same procedure is followed iteratively. It is a type of unsupervised learning algorithm [9]. The researchers Prashant and Govil in [10] examined the clustering analysis in data mining that analyzes the use of k-means algorithm in improving students academic performance in higher education and presents k-means clustering algorithm as a simple and efficient tool to monitor the progression of students performance.

#### Algorithm

1. Select the no. of 'c' cluster centers. (Fixed number of clusters is used in K-Means)
2. Initial cluster centers are determined for each of the c clusters, either by the software or by the researcher.
3. Find out the distance between each data point and cluster centers using Euclidean distance formula  $d(x, y) = \sqrt{\sum (x_i - y_i)^2}$ , where  $i = 1$  to  $n$ .
4. Assign the data point to that cluster center whose distance from the cluster center is minimum as compared to all the cluster centers.
5. Recalculate the new cluster center using

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_j$$

Where 'c<sub>i</sub>' represents the number of data points in i<sup>th</sup> cluster.

6. Recalculate the distances of each data point with new cluster centers.
7. If there is no reassigning of the data points, then stop, otherwise repeat step 3 onwards.

### 3.2 Hierarchical Clustering Algorithm

Hierarchical Clustering starts from 'n' number of clusters and ends up in a single cluster. Hierarchical Clustering is basically a set of nested clusters organized in a hierarchical tree [5].

Algorithm (Single-linkage cluster) [11]:

1. Assign a cluster to each item, such that N clusters for N items.
2. Find and merge the pair of clusters which are closest to each other.
3. Calculate the distances between the new and each of the old clusters. (using single-linkage cluster)
  - a) Start with the disjoint clustering having level  $l(0) = 0$  and sequence number  $n = 0$ .
  - b) In the current clustering, now find the least dissimilar pair of clusters say pair (a), (b), according to  $d[(a),(b)] = \min d[(u),(v)]$  where the minimum is over all pairs of clusters in the current clustering.
  - c) Increment the sequence number:  $n = n + 1$  and merge clusters (a) and (b) into a single cluster to form the next clustering n. Set the level of this clustering to  $l(n) = d[(a),(b)]$
  - d) Now the next step is to update the proximity matrix, M, by deleting the rows and columns corresponding to clusters (a) and (b) and adding a row and column correspond to the newly formed cluster. The proximity between the new cluster, denoted (a, b) and old cluster (k) is defined in this way:  $d[(k), (a, b)] = \min d[(k),(a)], d[(k),(b)]$
4. If all the objects are in one cluster then stop the process else, go to step 2.

### 3.3 BIRCH

The Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) is an agglomerative hierarchical clustering algorithm that performs well on large datasets with high dimensionality [12]. BIRCH creates a height-balanced clustering

feature tree of nodes that summarizes data by accumulating its zero, first, and second moments of the cluster. The resulting clustering feature (CF) is used to compute the centroids and measures the compactness and distance of the cluster.

### 3.4 DBSCAN

The Density Based Spatial Clustering of Application with Noise (DBSCAN) is a density-based technique that separates data points into three parts: Core points (points that are at the interior of the cluster), Border points (points which fall within the neighborhood of the core point) and Noise points (points that are not a core point or a border point). It utilizes a defined minimum radius ( $\epsilon$ ) and minimum cluster size to determine which density areas within the space can be considered a cluster [3].

## 4. CLUSTERING IN EDM

As a matter of fact, many people working in academia have started extensively applying Data mining techniques in the exploration for a better understanding of the student academic performance and their behaviors. There exists a rich body of literature dealing with the application of data mining in this novel perspective sometimes referred to as Educational data mining [13,14]. Cluster analysis used to separate a large set of data into subset. Clustering is the fundamental techniques used in analyzing the student data set. Moreover, this study makes use of cluster analysis to group student according to their features. This section describes literature survey done by other researchers using clustering in educational data mining.

Durairaj and Vijitha had developed a trust model using data mining techniques which mines required information for the education system to adopt this

strategic management tool [15]. A Research work by Oyelade [16], had presented a method of using K-means, clustering algorithm for the prediction of student's academic performance. The ability to monitor the progress of student's academic performance is a critical issue to the academic community of higher learning. The researcher aims to present a systematic review on different clustering techniques applied for educational data mining to predict academic performance of students and its implications. Another research by Shiwani and Roopali [17], had proposed a work to evaluate the performance of students of Digital Electronics of university institute of engineering and technology. The researcher had applied unsupervised learning algorithms such as K-means and Hierarchical clustering using WEKA tool as an open source tool.

A work by Azhar Ranf et.al had proposed a method known as K-means clustering, it calculated initial centroids instead of random selection, due to which the number of iterations is reduced and elapsed time is improved [18]. A Research work done by Veeramuthu et al. [19], had designed a model to present as a guideline for higher educational system to improve their decision-making processes. The authors aim to analyse how different factor affect a student learning behavior and performance using K-means clustering algorithm. A work done by Sivaram and Ramar [20] had surveyed the applicability of clustering and classification algorithms for recruitment data mining techniques that fit the problems which are determined. A study has been made by applying K-means, fuzzy C-means clustering and decision tree classification algorithms to the recruitment data of an industry.

## 5. CONCLUSION

The research paper has put an effort to reveal that the clustering techniques serve as powerful tool in educational data mining. Here various clustering algorithms are discussed and by using these algorithms student's performance is evaluated. The survey of clustering shows the majority of research paper used K-means clustering algorithm for evaluation and exploration. Furthermore, this simple review work describes that making use of cluster analysis to group student according to their features which can be used for the effective and faster results prediction by the educational institution. By concluding, his research work is presented as a guideline for the educational system to enhance their decision-making processes. With that, the suitable choice of clustering algorithm that justifies the research questions on student's data can be more effectively used for performance analysis from the large data set.

## REFERENCES

- [1] Romera, C., Ventura, S., Garcia, E., "Data mining in course management systems: Moodle case study and tutorial, Computer & Education, Vol.51, No. 1, 2008, pp.368-384.
- [2] Han, J., Kamber, M., Pei, J., "Data Mining: Concepts and Techniques", Elsevier, 2012.
- [3] Kyle DeFreitas Margaret Bernard, "Comparative performance analysis of clustering techniques in educational data mining", Int. Journal on Computer Science and Information Systems, Vol.10, No. 2, pp.65-78.
- [4] Sivogolovko, E., Novikov, B., "Validating cluster structures in data mining tasks", Proceeding of 2012 Joint EDBT/ICDT Workshops on EDBT-ICDT'12, USA: ACM, 2012, pp.245-250.
- [5] Mark Hall, et al., "The WEKA Data Mining Software: An update", SIGKDD

- Explorations Newsletter, Vol.11, No.1, 2009, pp.10-18.
- [6] Bijuraj L.V., “Clustering and its Applications”, National Conference on New Horizons in IT, 2013.
- [7] Berkhin, P.P., “A Survey of Clustering Data Mining Techniques”, Grouping Multidimensional data, Springer Berlin Heidelberg, 2006, pp.25-71.
- [8] Hämmäläinen, Wilhelmiina, Ville Kumpulainen, Maxim Mozgovoy, "Evaluation of clustering methods for adaptive learning systems", Artificial Intelligence Applications in Distance Education, 2013, pp. 1-32.
- [9] Khan, Dost Muhammad, Nawaz Mohamudally, "An agent oriented approach for implementation of the range method of initial centroids in k-means Clustering Data Mining Algorithm", REASON, Vol. 1, No.1, 2010, pp. 104.
- [10] Prashant Sahai Saxena, Govil M.C.,”Prediction of student academix performance using clustering”,IFRSA Int. J Data Warehouse Min, Vol.4, No. 2, 2014, pp.1-6.
- [11] Punitha, S. C., P. Ranjith Jeba Thangaiah, Punithavalli M., "Performance Analysis of Clustering using Partitioning and Hierarchical Clustering Techniques," Int. Journal of Database Theory and Application, Vol.7, No. 6, 2014, pp. 233-240.
- [12] Witten, I.H., Frank, E. Hall, M.A., “Data Mining: Practical Machine Learning Tools and Techniques”, Morgan Kaufmann, 2016.
- [13] Baker R., Yacef K., “The State of Educational Data Mining in 2009: A Review and Future Visions”, Journal of Educational Data Mining, Vol. 1, Issue 11, pp. 3– 17.
- [14] Romero C., Ventur S., “Educational Data Mining: A Review of the State-of-the-Art”, IEEE Transaction on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Vol. 40, No.6, 2010, pp.601-618.
- [15] Durairaj M.,Vijitha C.,” Educational Data mining for prediction of student performance using clustering algorithms”, Int. Journal of Computer Science and Information Technologies,Vol.5, No.4,2014,pp.5987-5991.
- [17] Shiwani Rana, Roopali Garg,”Evaluation of Student’s Performance of an Institute using Clustering Algorithms”, Int. Journal of Applied Engineering Research, Vol.11,No.5,2016,pp.3605-3609.
- [18] Azhar Rauf, Sheeba, “Enhanced K-mean clustering Algorithm to reduce number of Iterations and Time complexity”, Middle-East Journal of Scientific Research, Vol.12, No.7, 2012, pp.959-963.
- [19] Veeramuthu P.,Periyasamy R.,Sugasini V., “Analysis of student result using Clustering Techniques”, Int. Journal of Computer Science and Information Technologies, Vol.5, No.4, 2014, pp. 5092-5094.
- [20] Sivaram N., Ramar K., “Applicability of Clustering and Classification Algorithms for Recruitment Data Mining”, Int. Journal of Computer Applications, Vol. 4, No.5, 2010.