Asian Journal of Computer Science and Technology
ISSN: 2249-0701 Vol.8 No.2, 2019, pp. 85-89

Grouping of E Learners Using Fuzzy K-Medoid Clustering

Vidyaathulasiraman[1], S. Anthony Philomen Raj[2] and A. George Louis Raja[3]
[1]Department of Computer Science, Government Arts & Science College for women, Tamil Nadu, India
[2]Research Scholar, Periyar University, Tamil Nadu, India
[3]Department of Master of Computer Applications, Sacred Heart College, Tamil Nadu, India
E-Mail: vidyaathulasi@gmail.com, philomen@shctpt.edu, george@shctpt.edu

***Abstract* - The process of clustering in the general perspective is limited to the grouping of data into clusters and finds its applications in the fields of information retrieval, text ranking and classification and more. The dimension of e-Learning is to improve learning with various tools and technologies. Grouping of learners based on their learning levels is found to improve the learning abilities. Scientific method to cluster the learners is not available in literature, which can further simplify the amalgamation of learning complemented through clustering. This paper is an attempt to examine the aspects of implementing clustering to group the learners according to their learning abilities.**
***Keywords*: E-Learning, Grouping of Learners, Clustering**

I. INTRODUCTION

Clustering is the method of grouping together the objects of the same category. Qualitatively, behaviourally, semantically, or contextually analogous objects can be grouped together. In other words, homogeneous objects are grouped in one cluster and heterogeneous objects are grouped in another cluster. Clustering technique is widely used in many applications such as image processing, pattern recognition, market research, and data analysis.

II. CLUSTERING METHODS

The clustering methods are the following, 1. Partitional Clustering, 2. Hierarchical Clustering, 3. Density Based Clustering, 4. Grid Density-based Clustering, 5. Model-based Clustering and 6. Constraint-based Clustering [3].

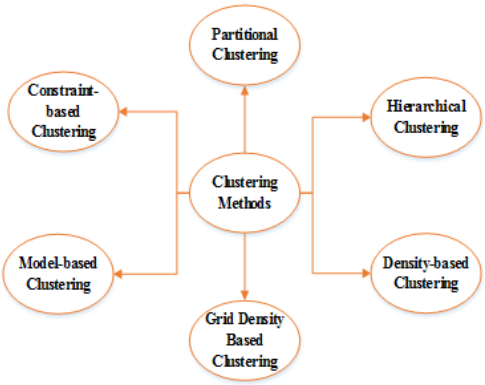

Fig. 1 Clustering Methods

A. Partitional Clustering: In the Partition Clustering each object belongs in exactly one cluster. This algorithm partitions the objects into k clusters and the number of clusters k is given well in advance. Partition clustering method groups the objects based on their nearest distance. In Partitioning methods, the cluster is mutually exclusive, and shape of the cluster is spherical. Mostly mean and median is used as a cluster centre to represent each cluster. This method is suitable for medium and small size of data set [8].

Pros
1. Scalable and Simple.
Cons
2. It is suitable only when the prior number of clusters is known.

B. Hierarchical Clustering: This clustering method considers a set of nested clusters that are represented as a tree. It produces a hierarchy of clusters called a dendrogram. This Hierarchical method produces hierarchical decomposition for the given dataset. Agglomerative and Divisive are the two types of Hierarchical Clustering**.** Another name for Agglomerative is “bottom-up approach”. In Bottom-Up approach the observation begins with many clusters and the clusters are combined based on the similarity of the objects. The hierarchy moves up every time the clusters are grouped together. The Divisive hierarchical clustering is also called “Top-Down” approach. The Top-Down approach uses the single cluster for the observation, and the fragmentation of the cluster is done recursively to narrow down the hierarchy. This type of clustering algorithm is mostly used for categorical data [1].

Pros
1. Easy to implement.
2. Good for small data sets.
Cons
1. Algorithm can never undo the previous step.
2. Not suitable for large datasets.

C. Density-Based Clustering: This clustering method which groups the data which is in the region with high density of the data space is considered to belong to the same cluster. The Density Connectivity and Density Reachability concepts are used in the Density based Clustering.

1. Density Connectivity: Points "p" and "q" are said to be density connected if there exists a point "r" which has

enough points in its neighbours and both the points "p" and "q" are within the ε distance. This is called chaining process. So, if "q" is neighbour of "r", "r" is neighbour of "s", "s" is neighbour of "t" which in turn is neighbour of "p" then it implies that "q" is neighbour of "p".

2. Density Reachability: A point "p" is said to be density reachable from a point "q" if point "p" is within ε distance from point "q" and "q" has enough points in its neighbours who are within distance ε.

Pros

a. Used for arbitrarily shaped clusters.
b. The algorithm is robust to outliers.

Cons

a. It is not completely determinable.
b. It is impossible to cluster different sets of data with great transformations in thicknesses.

D. Grid Density Based Clustering: Grid Density Based Clustering uses the densities of the neighbours to find the dense grids. The diverse shaped clusters are handled in the Grid Density Based Clustering for multi-density environs. The data grouped into one grid is called Grid density. The dense unit is defined when the grid density is higher than the density threshold value [6].

Pros

1. Uses a Multi-resolution grid data structure.
2. Has fast processing time.

Cons

1. The horizontal or vertical structures are used to form the clusters. So, no diagonal boundary is available.

E. Model-Based Clustering: The k-means and hierarchical clustering combine and form a heuristic approach to construct clusters. This clustering uses EM algorithm to identify the most likely model components and the number of clusters.

Pros

1. It provides estimates of the number of classes as well as their parameters.
2. It can directly compare the 'goodness' of clusters of different sizes.
3. The cluster definitions can overlap, which allows degree of 'fuzziness' for samples that are laid on the boundaries of different clusters.

Cons

1. In the view of large data set it takes too long to execute. So, it is suitable only for small size data set.

F. Constraint-Based Clustering: Different groups of data are categorized based on the user's preferences or constraints. This constraint-based clustering follows different approaches such as Constraint on individual objects, Obstacle objects as Constraints, Clustering parameter as Constraints, Constraints imposed on each individual cluster [7].

Pros

1. Based on the user or application the constraints are fixed.

III. E LEARNING

E-learning is defined as learning content or learning activities delivered using electronic tools which is also referred as e-content [9]. The process of e learning includes Virtual Classrooms, Digital Collaboration, Web Based Learning and Computer-Based learning. The e-content is delivered by different methodologies with the support of Internet/Intranet, audio or video, satellite TV, CD-ROM and any other LMS tools. Various LMS tools such as Moodle, Sakai are used to enhance the learning activities of e Learners [4].

IV. GROUPING OF LEARNERS

The purpose of grouping of leaners is to identify the learners' interest based on the frequency of their learning activities. In other words, it is used to select an appropriate learning activity which helps the learners to improve their learning skills in the process of teaching and learning. The most widely used learning activities are Q&A Sessions, Branching Scenarios, Serious Games, Real-World Examples, Video Demos, Online Discussions, eLearning Assessments, Simulations and Social Learning Opportunities [10]. Fig.2 portrays the architecture of grouping of learners.

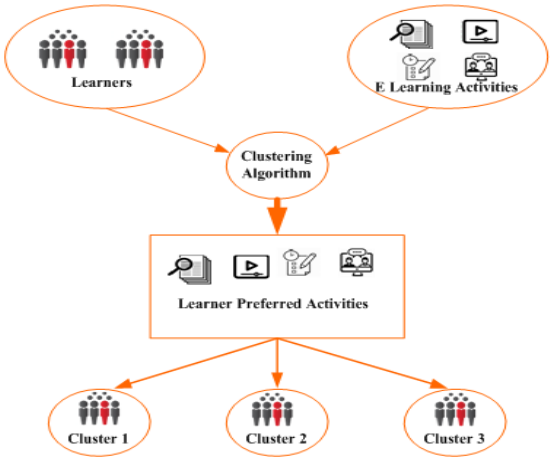

Fig. 2 Architecture of Grouping of Learners

Various learning activities are adopted by n number of learners. The preferred activities of the n number of learners are identified by applying the clustering algorithm. The algorithm used to categorize the learners is based on their preferred activities in different groups.

A. Benefits of Grouping of Learners

Teaching and Learning processes are important for both learners and teachers. The learning and teaching processes can be improved by identifying the learners' preferred e-learning activities by the teacher, so that the teacher can

implement the appropriate e-learning activities to the group of learners.

The grouping of learners is analogous to the process of clustering and classification. Since classification is a supervised mechanism for grouping, it incurs additional time complexity and space requirements and is suitable only for linear data, thus making its application inappropriate to group learners. This has led to the application of clustering to augment grouping of e Learners. The following section provides an extensive survey of various clustering strategies and implementation with detailed discussion and suggests the optimal clustering algorithm to group the learners.

V. PROSPECTS OF CLUSTERING LEARNERS

The survey of various clustering strategies has opened the prospective of applying clustering process to group the learners. However, the suitable method to be deployed to group the learners appropriately should be identified. The table I compares the various clustering methods based on eleven basic parameters and suggests the suitable clustering method to complement grouping of learners. The parameters are Methodology, Notation, Shape, Space Complexity, Time Complexity, Data Set Size, Data Set Type, Scalability, Multidimensional Data, Data Orientation, Data and Fuzziness [5].

1. *Methodology*: Describes the working methodology of a Clustering algorithm.
2. *Notation*: The mathematical notation of a clustering algorithm is denoted with the notation parameter.
3. *Shape*: The shape of clusters is verified.
4. *Space Complexity*: This is to measure the storage space of a clustering algorithm
5. *Time Complexity*: This is identifying the amount of time taken by a clustering algorithm.
6. *Data Set Size*: This parameter is to identify the data set size either small or large.
7. *Data Set Type*: This parameter is to identify the data set type either numerical or categorical data.
8. *Scalability*: This parameter is to identify the originality of a cluster both smaller and larger data set.
9. *Multidimensional Data*: This parameter is used to check whether the algorithm supports the Multidimensional Data or not.
10. *Data Orientation*: This parameter identifies the data type of an algorithm either Linear or Non-Linear data.
11. *Fuzziness*: This parameter is used to identify the fuzziness of an algorithm.

VI. COMPARISON TABLE FOR PARTITION CLUSTERING ALGORITHM

TABLE I COMPARISON TABLE FOR PARTITION CLUSTERING ALGORITHM [2]

S. No	Partitioning Clustering	Methodology	Notation	Shape	Space Complexity	Time Complexity	Data Set Size	Data Set Type	Scalability	Multidimensional Data	Data Orientation	Fuzziness
Linear Partitioning Clustering Algorithms												
1	K-Mean	The algorithm identifies the groups in the data based on the variable k. Where k is the number of groups.	$J(V)=\sum_{i=1}^{c}\sum_{j=1}^{c_i}(\|x_i-v_j\|)^2$	Spherical	O((m+k)n	O(n k d)	L	N	Yes	Yes	Linear	No
2	K-Medoid	The algorithm discovers the different groups of data by using midpoint data item from the data set.	$j=\sum_{i=1}^{k}\sum_{p\Omega c_j}^{n}\left\|P-O_j\right\|$	Arbitrary	O(ik(n-k)2).	$O(n^2dt)$	L	N/C	Yes	Yes	Linear	No
3	K-Modes	The goal of this algorithm is to use the categorical variable to find the different groups from the data set.	$d(X,Y)=\sum_{j=1}^{m}\delta(x_i,y_j)$	Arbitrary	O((n + **k**). ∑m.	O(n)	L	C	Yes	Yes	Linear	No
4	CLARA	The Clustering Large Applications (CLARA) uses the random of small size data from the actual data set. The different groups are deduced using PAM.	$Cost(M,D)=\frac{\sum_{i=1}^{n} dissimilarity(O_i,rep(M,O_i))}{n}$	Arbitrary	O(n)	$O(k(40+k)^2+k(n-k))$	L	N	Yes	Yes	Linear	No

5	CLARANS	Clustering Large Applications based on Randomized Search (CLARANS) uses the randomized search graph to categorize the different group of data.	$Cost(M,D) = \frac{\sum_{i=1}^{n} dissimilarity(O_i, rep(M,O_i))}{n}$	Arbitrary	O(n2)	O (kn^2)	L	N	Yes	Yes	Linear	No
6	PAM	The aim of this algorithm is to discover the different groups of data based on centrally located medoids.	$Z = \sum k\ i=1 \sum \|x\text{-}ci\|^2$	Spherical	O(k(n-k)^2)	$O(k(n-k)^2)$	S	N	No	Yes	Linear	No
Non-Linear Fuzzy Partitioning Clustering Algorithms												
1	K-Mean	The algorithm identifies the groups in the data based on the variable k. Where k is the number of groups.	$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2$	Spherical	O((m+k)n	O(n k d)	L	N	Yes	Yes	Non-Linear	Yes
2	K-Medoid	The algorithm discovers the different groups of data by using midpoint data item from the data set.	$j = \sum_{i=1}^{k} \sum_{p \Omega c_j}^{n} \left\| P - O_j \right\|$	Arbitrary	O(ik(n-k)2).	$O(n^2dt)$	L	N/C	Yes	Yes	Non-Linear	Yes
3	K-Modes	The goal of this algorithm is to use the categorical variable to find the different groups from the data set.	$d(X,Y) = \sum_{j=1}^{m} \delta(x_i, y_j)$	Arbitrary	O((n + **k**). ∑m.	O(n)	L	C	Yes	Yes	Non-Linear	Yes
4	CLARA	The Clustering Large Applications (CLARA) uses the random of small size data from the actual data set. The different groups are deduced using PAM.	$Cost(M,D) = \frac{\sum_{i=1}^{n} dissimilarity(O_i, rep(M,O_i))}{n}$	Arbitrary	O(n)	$O(k(40+k)^2 + k(n-k))$	L	N	Yes	Yes	Non-Linear	Yes
5	CLARANS	Clustering Large Applications based on Randomized Search (CLARANS) uses the randomized search graph to categorize the different group of data.	$Cost(M,D) = \frac{\sum_{i=1}^{n} dissimilarity(O_i, rep(M,O_i))}{n}$	Arbitrary	O(n2)	$O(kn^2)$	L	N	Yes	Yes	Non-Linear	Yes
6	PAM	The aim of this algorithm is to discover the different groups of data based on centrally located medoids.	$Z = \sum k\ i=1 \sum \|x\text{-}ci\|^2$	Spherical	O(k(n-k)^2)	$O(k(n-k)^2)$	S	N	No	Yes	Non-Linear	Yes

S = Small, L = Large, N = Numerical Data, C = Categorical Data

VII. PROPERTIES OF LEARNER GROUPS

The observations on the manual grouping of learners based on their learning activities into groups have conveyed the following inherent properties:

1. Learner Groups are Arbitrary Shaped: Learners for each learning activity may vary and the number of learners for each learning activity may vary.

2. Learner Groups are Formed Based on Numerical or Categorical Data: Learners are grouped based on their frequent access, performance and gender.

3. *Learner Groups are Mutually Inclusive:* Learner is grouped based on their preferred activities including other activities.

4. Learner Groups are Infinite in Size: The number of learners is not fixed under any circumstances.

VIII. RESULTS AND DISSCUSSION

The grouping of learners can be done using either clustering or classification processes. Classification is a supervised grouping mechanism involving training data and incurs additional time complexity compared to clustering methods. Hence, this paper has deployed the clustering mechanism to group learners.

From the table I, it can be inferred that, fuzzy k-medoid clustering better suits the grouping of learners. This is because it naturally satisfies the properties of learner groups as mentioned above. It is able to produce arbitrary shaped clusters, and can work with both numerical and categorical data, and can formulate clusters with overlapping objects which are mutually inclusive and can work for an unpredictable size of data. Confirming the above properties, we conclude that fuzzy k-medoid clustering can be applied for learner grouping.

IX. CONCLUSION

This article deliberates the architecture of Grouping of learners with the different clustering methods. The partition clustering method algorithms are identified to be appropriate and compared based on eleven different parameters. The four inherent properties of grouping the learners are described. It is proposed that the Fuzzy k-Medoid clustering better suits the grouping of e-Learners.

REFERENCES

[1] Dongkuan Xu and Yingjie Tian, "A Comprehensive Survey of Clustering Algorithms", *Ann. Data. Sci*. Vol. 2, No. 2, pp. 165–193, DOI 10.1007/s40745-015-0040-1, 2015

[2] T. Sajana, C. M. Sheela Rani and K. V. Narayana, "A Survey on Clustering Techniques for Big Data Mining", *Indian Journal of Science and Technology*, DOI:10.17485/ijst/2016/v9i3/75971, January 2016.

[3] Akshay S. Agrawal and Sachin Bojewar, "Comparative Study of various Clustering Techniques", *International Journal of Computer Science and Mobile Computing*, Vol. 3, No.10, pp. 497 – 504, October 2014.

[4] Kalpit G. Soni and Atul Patel, "Comparative Analysis of K-means and K-medoids Algorithm on IRIS Data", *International Journal of Computational Intelligence Research*, Vol. 13, No. 5, pp. 899-906, 2017.

[5] Ravi Sankar Sangam and Hari Om, "The k-modes algorithm with entropy based similarity coefficient", *2nd International Symposium on Big Data and Cloud Computing (ISBCC'15), Procedia Computer Science*, *50*, pp. 93 – 98, 2015.

[6] Jyotismita Goswami, "A Comparative Study on Clustering and Classification Algorithms", *International Journal of Scientific Engineering and Applied Science (IJSEAS)*, Vol. 1, No. 3, June 2015.

[7] Anthony K. H. Tung, Jiawei Han, Laks V. S. Lakshmanan and Raymond T. Ng, "Constraint-Based Clustering in Large Databases".

[8] K. Aparna and Mydhili K Nair, "A Detailed Study and Analysis of different Partitional Data Clustering Techniques", *International Journal of Innovative Research in Science, Engineering and Technology*, Vol. 3, No. 1, January 2014.

[9] V. Venkateswaran, R. Ramakrishnan and A. Ramalingam, "E-Learning Retrieval System Through Advanced Data Mining Clustering Technique", *International Journal of Modern Engineering Research (IJMER)*, Vol. 2, No. 4, pp. 1572-1575, July-Aug 2012.

[10] Herlina Latipa Sari, Dewi Suranti and Leni Natalia Zulita, "Implementation of k-means Clustering method for Electronic Learning Model", *International Conference on Information and Communication Technology (IconICT), IOP Conf. Series: Journal of Physics: Conf. Series*, Vol. 930, 012021, doi :10.1088/1742-6596/930/1/012021, 2017.