

GLOBAL ACADEMIC RESEARCH INSTITUTE

COLOMBO, SRI LANKA



GARI International Journal of Multidisciplinary Research

ISSN 2659-2193

Volume: 07 | Issue: 01

On 28th February 2021

<http://www.research.lk>

Author: B. Venura Lakshman, Dr. Janaka Wijayakulasooriya, Dr. Manjula Sandirigama

Bhiksu University, University of Peradeniya, Sri Lanka

GARI Publisher | Education | Volume: 07 | Issue: 01

Article ID: IN/GARI/ICCAET/2020/104 | Pages: 119-122 (04)

ISSN 2659-2193 | Edit: GARI Editorial Team

Received: 04.12.2020 | Publish: 28.02.2021

STUDENTS' DATA CLUSTERING IN EDUCATION SCIENCE

¹B. Venura Lakshman, ²Dr. Janaka Wijayakulasooriya, ³Dr. Manjula Sandirigama

¹Department of Practical Buddhist Studies, ²Department of Electrical and Electronic Engineering, ³Department of Computer Engineering,
¹Faculty of Buddhist Studies, ^{2,3}Faculty of Engineering,
¹Bhiksu University, ^{2,3}University of Peradeniya, Sri Lanka

ABSTRACT

Modern education uses many technological applications like Information Technology. In modern educational environments which are empowered by Information Technology generate huge data sets. Such data sets contain millions of individual records. With the explosive growth of the educational data sets, it is required to utilize sophisticated analysis tools and methods to work with these mega data sets. Therefore, various computer based clustering algorithms play an important role in such analysis. But in a computer system, it is better to perform the analysis with minimum human interactions in order to get results faster. So, in such cases, unsupervised clustering algorithms are very important as they lead the system to be autonomous. This paper describes how to use a newly designed unsupervised clustering algorithm to cluster students' marks for the best number of clusters. The newly designed unsupervised clustering algorithm estimates the best suited number of clusters for the data set and cluster automatically the data set. Also, the new algorithm was tested using a real data set which contains students' examination marks and results are also shown in the paper.

Keywords: Modern education, Educational data sets, Unsupervised clustering algorithm

CLUSTERING

What is clustering?

In clustering, (or cluster Analysis) similar objects in a data sets are grouped into a same group. (or a cluster).

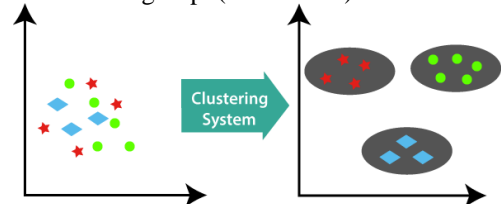


Fig 01: Clustering

Clustering are used many industries. For instance, Marketing segmentation, Social Network Analysis, image Segmentation...etc.

Classification of Clustering

There are various classifications in clustering. For instance, density based clustering, Distribution based clustering...etc.

Also all clustering algorithms can be classified as supervised and unsupervised. In supervised clustering, user must aware about number of clusters which is best suited to the data set. But in unsupervised clustering the algorithm autonomously decide the best number of clusters for the given data set.

Developing a new Unsupervised Clustering Algorithm

The prime objective of clustering is, to segregate groups with similar traits. There are various types of clustering algorithms. For the purpose of clustering a feature vector consisting of selected evaluation parameters are used. Usually, clustering algorithms deal with a single process. However, in Teaching and Learning Process, there are two separate processes which have to be considered.

1. The learning process basically depended on student's ability. In addition, the response time (RT) and guessing factor are also affected.

2. Usually, the teacher started to teach any lesson from the basic level. Therefore, the difficulty level of the lesson is started from the bottom and extended to the high level of difficulty.

But in study about both supervised and unsupervised algorithms, there should be a mechanism to find the next level of the cluster point. Thus, for the new algorithm, there should be a way of advancing to the next clustering level. As it is a psychometric measurement, Greatest Common Divisor (GCD) can be utilized as a factor of the increment. Therefore, by using GCD, it can be found the next advancing cluster point.

Further, a basic structure is required for the new clustering algorithm. It has to be a two phase algorithm with GCD as a factor to find next cluster level.

The new algorithm does not need a specific number of clusters to be given, before performing the clustering process and it is able to find the optimal number of clusters during the clustering process. For this task, the natural phenomenon of understanding of human mind is used. The new algorithm has two phases as shown below.

Input: k: number of clusters (for dynamic clustering initialize $k=2$) Fixed number of clusters = yes or no (Boolean).

Output: A set of clusters.

Method:

Teacher Phase

The teacher is the person who presents lessons. He presents lessons from the easy level to hard level. So, when he presents a basic level lesson, almost all students can understand it. Thus, there are either no clusters or two clusters. After delivering many more hard lessons, we can clearly see students divided themselves according to their understanding of lessons. (in any examinations also).

1. Suppose that, the difficulty of the lesson presented by teacher T1 $\rightarrow b_1$

2. At the initial level at the mean ability of the class a_1 Then, $X_1 = a_1$

3. After presenting the lesson with difficulty b_1 by the teacher T1, the mean ability of the class has been changed into anew then;

The difference of the mean $D = (a_{new} - a_1)$

4 Suppose that student $X_{old,i}$ status is transformed to $X_{new,i}$ for $i=1$ to n

5. Therefore, the function of the transformation $X_{new, I} = X_{old, I} + D$

Student Phase

The Student learns from the teacher. But their ability is different from each other. Therefore, when the teacher presents lessons from lower difficulty level to higher difficulty level, we can see students dividing themselves into clusters according to the understanding of lessons. (truly related to student's ability.) Therefore, it can be assumed that clustering of students is directly related to their ability & difficulty of the lesson. Therefore, calculate the probability of the respective students can be calculated & cluster them probability ranges after delivering each lesson.

Suppose that teacher T_i teaches lessons (easy to hard) average difficulty of b_i the mean ability a_i is denoted the success of population (the class).

After the end of one lesson, the teacher moves to much harder one. Clusters are defined according to the range of probability measurements range (r_i). Then, within certain range of probability clusters can be defined. For r_i probability range C_i cluster is defined. Then, clusters are formed automatically based on probability ranges. Therefore, it is not required to pre-define number of clusters for the population. This algorithm automatically defines required best number of clusters for the population.

After calculating the probability of each item, it can be calculated the maximum & minimum probability of all calculations. Then, the Greatest Common Divisor (GCD) of all probability values can be calculated. Afterword, the system can calculate suitable probability ranges starting from the minimum probability (P_{min}) to Maximum Probability (P_{max}). Next, the clusters and relevant centroids are calculated for the data set.

Suppose that, C_1 & C_2 are cluster centers based in probability ranges. $X_{new,I}$ & $X_{new,j}$ are two students in the class.

Then,

From P_1 to P_n do , ($n \in N$)

3. $P_1 = P_{min}$

4. If $P(c_1) > X_{new,I}$ and if $P(c_1) < X_{new,j}$ then
5. $X_{new,j} \in \{C_1\}$
6. Else $X_{new,I} \in \{C_2\}$
7. Until $P_n = P_{max}$ Loop

Here, the new unsupervised clustering algorithm is shown to cluster students, according to their psychometric measurements. Actually, this is a two tier algorithm. Therefore, when working with

the algorithm, the programmer must make the connectivity with both phases.

Testing the new two phase clustering mechanism.

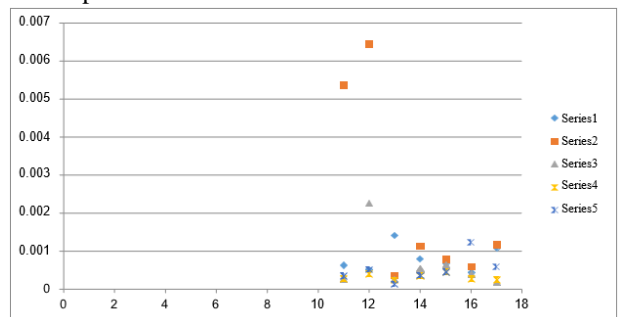
Source: MCQ paper with six questions (46 students)

To test the working level of the new two phase clustering algorithm, the real data are required. The same data set has been utilized for the clustering. Afterward, using Microsoft Excel, the new algorithm is tested and results are shown below.

By applying the algorithm, there are 14 clusters as shown below.

Fig 2.0(a): Cluster values after execution of new clustering algorithm.

Fig 2.0 (b): Cluster values after execution of new clustering algorithm – Graphical View.



Here, it is shown how the new unsupervised algorithm can be applied to a real data set and results are shown in 2.0 (a) and 2.0(b). It is clear that the algorithm is working without supervision and makes the best numbers of clusters for the given data set. There are cluster points which do not have any data points as this is a small data set. Usually, unsupervised clustering algorithms are working perfectly with

larger data sets. So, with a larger data set, there could be more data points.

REFERENCES

- C. Ching-Yi and Y. Fun, "Particle swarm optimization algorithm and its application to clustering analysis," in *IEEE International Conference on Networking, Sensing and Control*, 2004, vol. 2, pp. 789-794.
- S. Das, A. Abraham, and A. Konar, "Automatic clustering using an improved differential evolution algorithm," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 38, pp. 218-237, 2012.
- S.Z. Selim, K.S. Al-Sultan, "A simulated annealing algorithm for the clustering problem," *Pattern Recogn.* 24, 1003–1008.