Data Clustering in Education Science

B. Venura Lakshman

Introduction

National education is playing a key role in national development. Therefore, right decisions on national education are required to be taken to develop the future of a country. But with the explosive growth of the educational data sets and with challenging modern educational environment, it is required to utilize sophisticated tools and methods to take right decisions on national education.

As collections of students' data sets are huge, these data sets are required to be analyzed by using a computer system to get right decisions at the right time as quick as possible. Therefore, various computer based clustering algorithms play an important role in such analysis. But in a computer system, it is better to perform such analysis with minimum human interactions. So, in such cases, unsupervised algorithms are very important as they lead the system to be autonomous.

In this paper it is going to show how to use a new unsupervised clustering algorithm to cluster students' marks for the best number of clusters.

Clustering.

What is clustering?

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

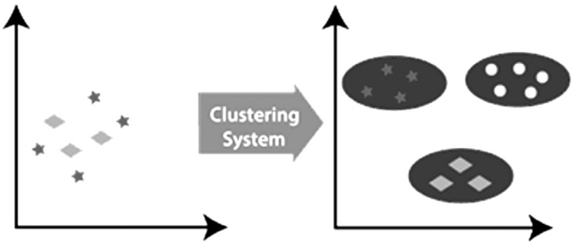


Fig 01: Clustering

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

2.3 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.

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

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).

- Suppose that, the difficulty of the lesson presented by teacher T_1 –>b_1
- At the initial level at the mean ability of the class a_1 Then, X_1= a_1
- After presenting the lesson with difficulty b_1 by the teacher T_1, the mean ability of the class has been changed into a_{new} then;
- The difference of the mean D= (a_{new}-a_1)
- Suppose that student $X_{old,i}$ status is transformed to $X_{new,i}$ for i=1 to n
- 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 Divider (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.

1. 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,

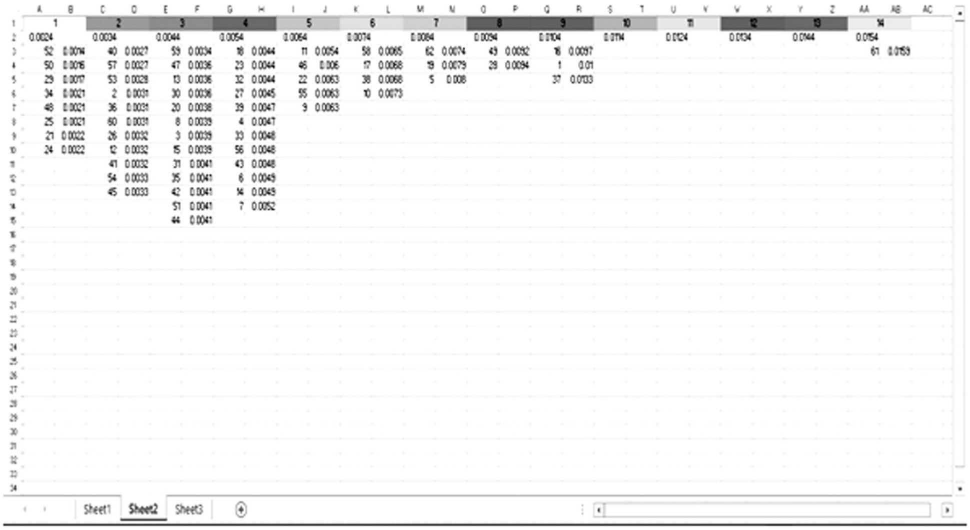
2. From P1 to Pn do , ($n \in N$)
3. $P1= P_{min}$
4. If $P(c1) > X_{new,I}$ and if $P(_{c1}) < X_{new,j}$ then
5. $X_{new,j} \in \{C1\}$
6. Else $X_{new,I} \in \{C2\}$
7. Until $P_n = P_{max}$ Loop

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.

3.1 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.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1	1		2		3		4		5		6		7		8		9		10		11		12		13		14	
2	0.0024		0.0034		0.0044		0.0054		0.0064		0.0074		0.0084		0.0094		0.0104		0.0114		0.0124		0.0134		0.0144		0.0154	
3	52	0.0014	40	0.0027	59	0.0034	18	0.0044	11	0.0054	58	0.0065	62	0.0074	49	0.0092	16	0.0097									61	0.0153
4	50	0.0016	57	0.0027	47	0.0036	23	0.0044	46	0.006	17	0.0068	19	0.0079	28	0.0094	1	0.01										
5	29	0.0017	53	0.0028	13	0.0036	32	0.0044	22	0.0063	38	0.0068	5	0.008			37	0.0103										
6	34	0.0021	2	0.0031	30	0.0036	27	0.0045	55	0.0063	10	0.0073																
7	48	0.0021	36	0.0031	20	0.0038	39	0.0047	9	0.0063																		
8	25	0.0021	60	0.0031	8	0.0039	4	0.0047																				
9	21	0.0022	26	0.0032	3	0.0039	33	0.0048																				
10	24	0.0022	12	0.0032	15	0.0039	56	0.0048																				
11			41	0.0032	31	0.0041	43	0.0048																				
12			54	0.0033	35	0.0041	6	0.0049																				
13			45	0.0033	42	0.0041	14	0.0049																				
14					51	0.0041	7	0.0052																				
15					44	0.0041																						

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

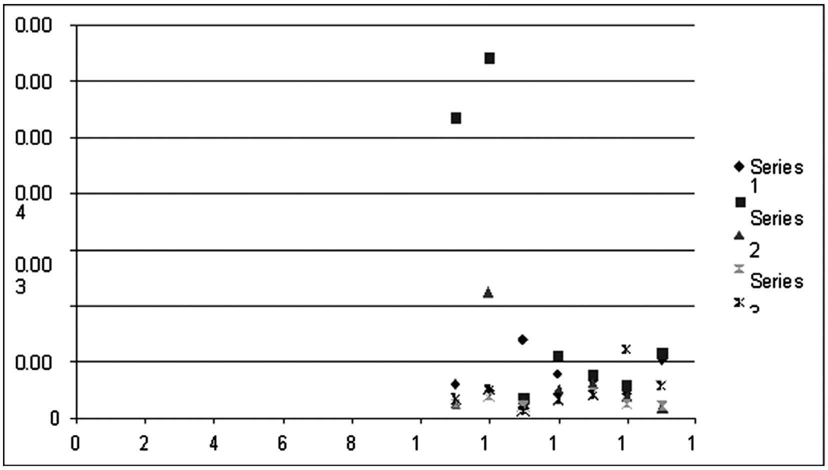


Fig 2.0 (b): Cluster values after execution of new clustering algorithm – Graphical View.

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.