Gustafson-Kessel Clustering Model To Identify The High-Level Performers In Educational Domain

S.Hamsanandhini, T.Thilagaraj, Dr.N Sengottaiyan

Abstract: The Process of placing individual items into groups based on some quantitative information is known as clustering. The cluster number estimation is one of the most important tasks in clustering. The Gustafson-Kessel clustering model will suitable to evolve many real-world tasks. This model having the ability to deal with unlabeled data and also it will generate membership and typicality values. The skill sets like speaking, reading, listening and writing are used to analyze the student level. The prediction of the best performer will reduce the burden in placement and also reduce the cost in the student enrichment process.

Index Terms: Data mining, Gustafson-Kessel Clustering, Placement, Student Performance

1. INTRODUCTION

In data mining, the primary objective is the prediction and description of statistics [1]. The data mining techniques like fuzzy, k-means, neural and decision trees are suitable to extract the hidden knowledge in large repository [2]. A similar group of an object is separated in clustering is a type of data model that shows historical perceptions [3]. The representation of elements in more than one group will be defined as fuzzy clustering [4]. The discovery of meaningful groups within a particular set of data items or vectors is mentioned as clustering. Different techniques are available to represent measuring between all elements of data in clustering. To find the cluster radius the median operation is used [5]. The vector concept is more suitable to store a large volume of data which can be very easy to process in R-Tool [6]. The valid predictions are made by pulling some useful facts in a huge repository [7]. The constructed classifier will more useful for predictions in the existing area with better accuracy [8]. The prediction process in the early stage will help the teachers to identify the various levels of students to assist proper training for their carrier [9]. The number of factors is suggested by the researchers to analyze student performance. These factors will use in the time of pre-placement training to identify the low and high-level performers [10]. In a few clustering processes, the fuzzy models will not provide sharp boundaries. Instead of crisp assignments zero and one are placed in fuzzy clustering models [11].

2 RELATED WORK

Sarbu, et al. [12] Here the field of clustering analysis is huge which holds many algorithms like FCM, GK, FPCM, MPFCM and more to deal with a variety of problems. The comparative study has made between fuzzy models while dealing with soil samples. In that Gustafson-Kessel clustering algorithm produced good results while comparing it with all others. The fuzzy set is a generalization of normal set theory which designed for partial truth concepts. The hierarchical clustering algorithm and Gustafson-Kessel cluster algorithm are majorly compared using the soil dataset. The fuzzy set shows ill-defined objects with improper boundaries in membership and non-membership subsets.

Egrioglu, et al. [13] In this paper the forecasting performance was analyzed using the Gustafson-Kessel algorithm to obtain the best Mean Squared Error. The representations of discrete fuzzy sets in fuzzy time series are different from discrete fuzzy sets in conventional time-series approaches. The increase of interests with researchers in fuzzy models implementations in difficult situations. The improvement process is happening in fuzzification and defuzzification stages using various research tools. Lesot and Kruse [14] Here the extensions of typicality degree to form clusters have made. In supervised learning, the typicality degree is the tool to build data categories using characteristic representatives. The internal resemblance and another important fact external dissimilarity are taken into account in typicality degrees. Teslic, et al. [15] The GK clustering is used to find the cluster center's initial location and it works with the combination of local model network learning procedure. The cluster was split into two cluster's and two cluster centers are formed to make the clustering task more effective. Keller [16] Here the approach has made to make distance measure is varied it can be considered as a generalization of fcm and Gustafson-Kessel algorithm. The new factor introduced for individual data points that will reduce the outlier influence in the entire classification process. Identification of critical data that enables us to split outliers for an entire dataset for an expert's analysis.

3 GUSTAFSON-KESSEL CLUSTERING ALGORITHM

The Gustafson Kessel Algorithm uses adaptive distance measures to find different sizes of clusters. The objective function of the GK Clustering algorithm to find different clusters (1). To find the Mahalanobis distance the data object is M and the cluster prototype is N (2). The matrix P holds n length tuples of the specific cluster (3).

$$J_{GK}(M,N,P,Q) = \sum_{i=1}^{nt}\sum_{j=1}^{mt} q_{ij}^{r}\, dP_j\left(\vec{m}_i,\vec{n}_j\right) \qquad (1)$$

$$dP_j\left(\vec{m}_i,\vec{n}_j\right) = \left(\vec{m}_i - \vec{n}_j\right)^T P_j \left(\vec{m}_i - \vec{n}_j\right) \qquad (2)$$

$$P = \{P_1, P_2, P_3, \ldots., P_n\} \qquad (3)$$

- *S.Hamsanandhini – Department of Computer Technology-PG, Kongu Engineering College, Perundurai, Tamil Nadu, India, hamsanandhini@gmail.com*
- *T.Thilagaraj – Department of Computer Applications, Kongu Arts and Science College, Erode -638107, Tamil Nadu, India, thilagaraj.t@gmail.com*
- *Dr.N.Sengottaiyan, Director, Sri Shanmugha College of Engineering and Technology, Sankari - 637304 , Tamil Nadu, India, sriram3999@gmail.com*

The Gustafson Kessel algorithm must satisfy the following conditions (4) and (5).

$$\sum_j^{mt} q_{ij}^r = 1; \quad 1 \le i \le nt \tag{4}$$

$$\sum_{i=1}^{nt} q_{ij}^r > 0; \quad 1 \le j \le mt \tag{5}$$

The Gustafson Kessel objective function is minimized by the updated equations (6) and (7).

$$q_{ij}^r = \left[\sum_{j=1}^{mt} \left(\frac{dP_j(\vec{m}_i, \vec{n}_j)}{dP_l(\vec{m}_i, \vec{n}_j)} \right)^{1/(r-1)} \right]^{-1} ; 1 \le i \le nt, 1 \le l \le mt \tag{6}$$

$$\vec{n}_j = \frac{\sum_{i=1}^{nt} q_{ij}^r \vec{m}_i}{\sum_{i=1}^{nt} q_{ij}^r}; \quad 1 \le j \le mt \tag{7}$$

TABLE 1: *DATA SET OF 30 STUDENTS FOR PLACEMENT ANALYSIS*

St.Id	Speaking	Reading	Writing	Listening
1	54	58	61	82
2	66	69	63	53
3	65	75	70	77
4	44	54	53	53
5	69	73	73	88
6	74	71	80	71
7	73	74	72	33
8	69	54	55	82
9	67	69	75	52
10	70	70	65	58
11	62	70	75	56
12	69	74	74	79
13	63	65	61	39
14	56	72	65	62
15	40	42	38	69
16	97	87	82	59
17	81	81	79	67
18	74	81	83	45
19	50	64	59	60
20	75	90	88	61
21	57	56	57	39
22	55	61	54	58
23	58	73	68	63
24	53	58	65	41
25	59	65	66	61
26	50	56	54	49
27	65	54	57	44
28	55	65	62	30
29	66	71	76	80
30	57	74	76	61

Table 1 shows the four attributes are speaking, reading, writing and listening skills of 30 students who have opted for placement. These are the essential skills needed for students and it will resemble their interpersonal.

TABLE 2*: THE FUZZY MEMBERSHIPS DEGREES OF DATA OBJECTS BY GUSTAFSON-KESSEL CLUSTERING ALGORITHM*

St. Id	Low (Cluster 4)	Medium (Cluster 3)	High (Cluster 2)	Very High (Cluster 1)
1	0.001946101	0.105893243	0.40291962	0.489241036
2	0.004844237	0.012919796	0.015450636	0.966785331
3	0.030217735	0.965353243	0.003150533	0.001278489
4	0.005890246	0.209673487	0.779546547	0.00488972
5	0.060909091	0.919188482	0.0158161	0.004086327
6	0.156734059	0.83094093	0.011350817	0.000974195
7	0.002790626	0.024763731	0.970797457	0.001648186
8	0.005486718	0.006160212	0.019084675	0.969268394
9	0.091980737	0.886271722	0.020906216	0.000841325
10	0.003430601	0.008082258	0.014502506	0.973984636
11	0.942299468	0.038135601	0.019264615	0.000300315
12	0.058694664	0.929598659	0.006190466	0.005516211
13	0.006721995	0.015278845	0.975409543	0.002589616
14	0.02241048	0.972340995	0.002708526	0.002539999
15	0.002313842	0.026742067	0.938100129	0.032843962
16	0.000642796	0.005112955	0.014551504	0.979692746
17	0.001529982	0.019734359	0.011161578	0.967574081
18	0.002280966	0.113129744	0.883542689	0.001046601
19	0.063394925	0.901828137	0.010620293	0.024156646
20	0.001576061	0.03876898	0.00490119	0.954753769
21	0.985615496	0.008060007	0.006013298	0.0003112
22	0.973009805	0.011567304	0.007651045	0.007771847
23	0.050908297	0.929393963	0.007534168	0.012163572
24	0.025018436	0.96635115	0.008141896	0.000488518
25	0.021578754	0.209872133	0.767371116	0.001177997
26	0.021642158	0.039775421	0.935608799	0.002973621
27	0.994700688	0.003036794	0.002037971	0.000224547
28	0.011352277	0.0712921	0.91583842	0.001517203
29	0.031825857	0.155742612	0.784797182	0.02763435
30	0.978783706	0.007773743	0.011557633	0.001884919

Table 2 shows the membership degrees for four ranges of clusters low, medium, high and very high using the Gustafson-Kessel clustering model. Table 3 shows clusters centers for low, medium, high and very high range with it respective categories. Table 4 shows descriptive statistics for membership degrees. The low performers are in cluster 4, medium in cluster 3, high in cluster 2 and very high in cluster 1.

TABLE 3: *THE CLUSTER CENTERS OF LOW, MEDIUM, HIGH AND VERY HIGH USING GUSTAFSON-KESSEL CLUSTERING ALGORITHM*

Levels	Speaking	Reading	Writing	Listening
Low (Cluster 4)	75.48523	74.42299	71.49527	64.09121
Medium (Cluster 3)	58.33294	63.44617	62.29151	50.09524
High (Cluster 2)	61.85708	69.74747	69.48658	65.63266
Very High (Cluster 1)	59.29688	62.90129	63.71448	51.60747

TABLE 4: *DESCRIPTIVE STATISTICS FOR THE MEMBERSHIP DEGREES BY CLUSTERS USING GUSTAFSON-KESSEL CLUSTERING ALGORITHM*

Description	Low (Cluster 4)	Medium (Cluster 3)	High (Cluster 2)	Very High (Cluster 1)
Size	7	9	9	5
Min	0.489241	0.7673711	0.8309409	0.9422995
Q1	0.9607696	0.7847972	0.9018281	0.9730098
Mean	0.9001857	0.8834458	0.922363	0.9748818
Median	0.9675741	0.9158384	0.929394	0.9787837
Q3	0.9716265	0.9381001	0.9653532	0.9856155
Max	0.9796927	0.9754095	0.972341	0.9947007

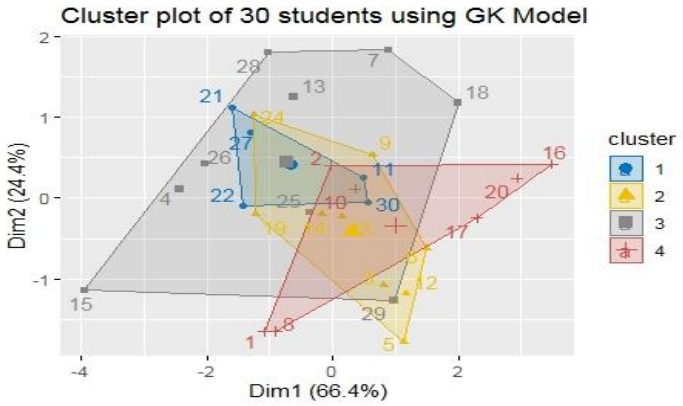
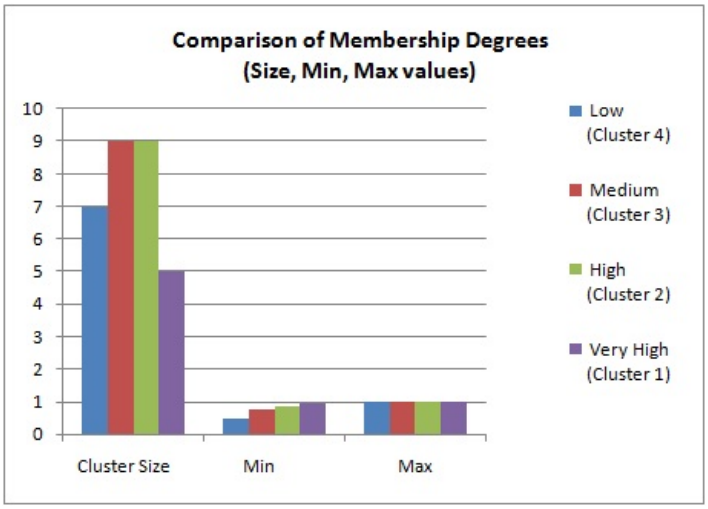


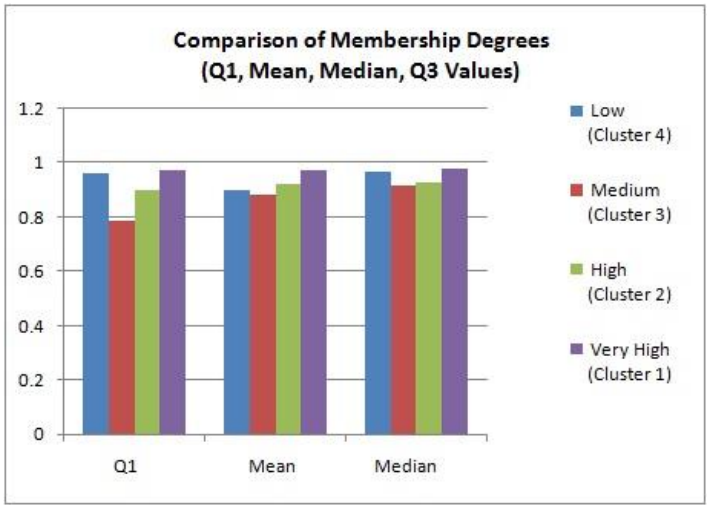
FIGURE 1: *RESULT OF THE FOUR CLUSTERS USING GUSTAFSON-KESSEL CLUSTERING ALGORITHM*

Figure 1 shows the result of the Gustafson-Kessel clustering model implemented using the R Tool. The four levels of the cluster are plotted which helps to identify the high-level performers.

GRAPH 1: *COMPARISON OF MEMBERSHIP DEGREES FOR SIZE, MINIMUM, AND MAXIMUM VALUES USING GUSTAFSON-KESSEL CLUSTERING ALGORITHM*



GRAPH 2*: COMPARISON OF MEMBERSHIP DEGREES FOR Q1, MEAN, MEDIAN, Q3 VALUES USING GUSTAFSON-KESSEL CLUSTERING ALGORITHM*



Graph 1 represents the cluster size, minimum, and maximum membership values. The comparison was made with different levels of clusters. Graph 2 represents Q1, Mean and Median values of membership degrees. These values are compared with all levels of clusters.

CONCLUSION

Here the Gustafson-Kessel algorithm is more efficient and easy to implement in difficult situations to find a different level of performers. This model will pick high performers and it will use to academia and recruiters at the time of placement. Early prediction of students regarding placement will reduce the cost. Among a different number of clusters, this model has high efficiency to generate high-level range cluster.

REFERENCES

[1] T. Silwattananusarn and K. Tuamsuk, "Data mining and its applications for knowledge management: a literature review from 2007 to 2012," arXiv preprint arXiv:1210.2872, 2012.

[2] B. K. Baradwaj and S. Pal, "Mining educational data to analyze students' performance," arXiv preprint arXiv:1201.3417, 2012.

[3] P. Berkhin, "A survey of clustering data mining techniques," in Grouping multidimensional data: Springer, 2006, pp. 25-71.

[4] A. Gosain and S. Dahiya, "Performance analysis of various fuzzy clustering algorithms: a review," Procedia Computer Science, vol. 79, pp. 100-111, 2016.

[5] L. Serir, E. Ramasso, and N. Zerhouni, "Evidential evolving Gustafson–Kessel algorithm for online data streams partitioning using belief function theory," International journal of approximate reasoning, vol. 53, no. 5, pp. 747-768, 2012.

[6] R. Pandey, N. Srivastava, and S. Fatima, "Extending R Boxplot Analysis to Big Data in Education," in 2015 Fifth International Conference on Communication Systems and Network Technologies, 2015: IEEE, pp. 1030-1033.

[7] B. M. Varghese, A. Unnikrishnan, G. Sciencist, N. Kochi, and C. Kochi, "Clustering student data to characterize performance patterns," Int. J. Adv. Comput. Sci. Appl, vol. 2, pp. 138-140, 2010.

[8] X. Wu et al., "Top 10 algorithms in data mining," Knowledge and information systems, vol. 14, no. 1, pp. 1-37, 2008.
[9] M. Gera and S. Goel, "A model for predicting the eligibility for placement of students using data mining technique," in International Conference on Computing, Communication & Automation, 2015: IEEE, pp. 114-117.
[10] P. S. Saxena and M. Govil, "Prediction of Student's Academic Performance using Clustering," in Natl. Conf. Cloud Comput. Big Data, 2009.
[11] D. Vanisri and C. Loganathan, "An efficient fuzzy clustering algorithm based on modified k-means," International Journal of Engineering Science and Technology, vol. 2, no. 10, pp. 5949-5958, 2010.
[12] C. Sarbu, K. Zehl, and J. W. Einax, "Fuzzy divisive hierarchical clustering of soil data using Gustafson–Kessel algorithm," Chemometrics and intelligent laboratory systems, vol. 86, no. 1, pp. 121-129, 2007.
[13] E. Egrioglu, C. Aladag, U. Yolcu, V. R. Uslu, and N. A. Erilli, "Fuzzy time series forecasting method based on Gustafson–Kessel fuzzy clustering," Expert Systems with Applications, vol. 38, no. 8, pp. 10355-10357, 2011.
[14] M.-J. Lesot and R. Kruse, "Gustafson-Kessel-like clustering algorithm based on typicality degrees," in Uncertainty And Intelligent Information Systems: World Scientific, 2008, pp. 117-130.
[15] L. Teslic, B. Hartmann, O. Nelles, and I. Skrjanc, "Nonlinear system identification by Gustafson–Kessel fuzzy clustering and supervised local model network learning for the drug absorption spectra process," IEEE transactions on neural networks, vol. 22, no. 12, pp. 1941-1951, 2011.
[16] A. Keller, "Fuzzy clustering with outliers," in PeachFuzz 2000. 19th International Conference of the North American Fuzzy Information Processing Society-NAFIPS (Cat. No. 00TH8500), 2000: IEEE, pp. 143-147.