Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

ISSN NO. 2320-5407

Article DOI: 10.21474/IJAR01/11171
DOI URL: http://dx.doi.org/10.21474/IJAR01/11171

RESEARCH ARTICLE

APPLICATION OF CLUSTERING TECHNIQUES TO STUDY THE TRAINING PATTERN PROVIDED BY THE DIFFERENT INSTITUTES UNDER HSRT

Ranjan Kumar Gupta[1], Sudatta Banerjee[2] and Debdip Khan[3]

1. Assistant Professor in Management, Department of Commerce and Management, West Bengal State University, Barasat, W.B., India.
2. Faculty, University Institute of Technology, Burdwan, W.B., India.
3. Faculty, Department of Business Administration, Burdwan Raj College, Burdwan, W.B., India.

Manuscript Info

Manuscript History
Received: 10 April 2020
Final Accepted: 12 May 2020
Published: June 2020

Key words:-
HSRT, Agglomerative Clustering Techniques, K-Mean Clustering Technique, Training Institutes

Abstract

In Social science researches of modern times, clustering techniques play an important role. Some common broad areas and also some specialized areas in which clustering has been successfully implemented are Science and Technology, Social Sciences, Humanities, Engineering, Medical Science, Data mining, Machine Learning, Pattern recognition, Image analysis, Information retrieval and Management. In this paper an effort has been made to apply common clustering techniques like Hierarchical Agglomerative Clustering and K-mean clustering in analysing the training types and patterns of different training institutes under HSRT i.e., “Hunar Se Rozgar Tak”, a special initiative of Ministry of Tourism, Government of India. Secondary data related to four different courses offered under HSRT by different training institutes of India has been collected from progress report for HSRT. The said data has been subjected to different clustering techniques with the view to reveal and analyze various underlying facts like similarity or differences between the training patterns, the role of time in changing the training pattern, impact of Geographical position, modification and alteration of training infrastructure etc.

Introduction:-

To serve different important purposes, grouping of objects has become a crucial task in various fields like Science and Technology, Social Sciences, Humanities, Engineering, Medical Science, Data mining, Machine Learning, Pattern recognition, Image analysis, Information retrieval and Management (Saxena et al., 2017; Ghuman, 2016). Clustering may be understood as a process of grouping a set of objects in a manner, so that the members of the same groups are closer/ similar, on the basis of some criterion, to each other than to the members of different groups.

Depending on the objective of the study, type of the objects and the type of data available, different clustering methods have been proposed in the literature. Reason for having different clustering approaches towards various techniques is due to the fact that there is no such precise definition to the notion of “Cluster” (Rokach, 2005; Castro and Yang, 2000). Clustering technique according to Fraley and Raftery (1998) can be broadly classified into two groups: Hierarchical Clustering Techniques and Partitional Clustering Techniques. This apart, there are other

Corresponding Author:- Ranjan Kumar Gupta
Address:- Assistant Professor in Management, Department of Commerce and Management, West Bengal State University, Barasat, W.B., India.

categorization of clustering techniques like density based methods, model based methods and grid based methods, as suggested by Han et al.(2011). One non-hierarchical simple clustering technique is the K-mean clustering technique as proposed by MacQueen (1967). Hierarchical Clustering can further be subdivided into Agglomerative and Divisive Clustering. Both these methods can be further grouped into three categories: Single linkage clustering, complete linkage clustering and average linkage clustering. In this work we have employed Agglomerative single linkage clustering and K-mean Clustering.

Though clustering techniques were known form the early part of eighteenth century but the major applications of that were started in the second half of nineteenth century. Ward (1963) proposed Hierarchical grouping to optimize an objective function. In 1967, MacQueen suggested some methods for classification and Analysis of Multivariate Observations and King proposed Step-wise Clustering Procedures. After that Zahn (1971) tried Graph-theoretical methods for detecting and describing gestalt clusters. A Fuzzy Relative of the ISODATA Process and its use in Detecting Compact Well-Separated Clusters was addressed by Dunn in 1973. In the same year Sneath and Sokal focused on Numerical Taxonomy. Urquhart (1982) introduced Graph-theoretical clustering based on limited neighbourhood sets. A survey of recent advances in hierarchical clustering algorithms which use cluster centers was also done (Murtagh, 1984).Gath and Geva (1989) worked on optimal fuzzy clustering and Conceptual clustering, categorization and polymorphy were taken care by Hanson and Bauer (1989).

In 1992, Celeux and Govaert classified EM algorithm for clustering on the basis of two stochastic versions. Krishnapuram and Keller (1993) addressed a probabilistic approach to clustering. Wolpert and Macready (1997) dealt with the Theorem for Optimization. Next Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications was addressed by Agrawal et al.(1998). Nakayama and Kagaku (1998) classified Pattern by linear goal programming and its extensions. Jain et al. (1999) reviewed Data Clustering and in 2000, Sheikholeslami et al. addressed Wave Cluster. Dolnicar (2003) Used Cluster Analysis for Market Segmentation for typical misconceptions, established methodological weaknesses and made some recommendations for improvement.

Law et al. (2004) introduced Multi objective Data Clustering and Xu and Wunsch (2005) made survey of clustering algorithms. In 2006, Faceili et al. worked on Multiobjective Clustering ensemble. Handl and Knowles (2007) proposed an evolutionary approach to Multiobjective clustering. In the same year Luxburg gave a tutorial on spectral clustering. Brendan and Dueck (2007) worked on Clustering by passing messages between data points.

In the year 2010, different types of researches were done such as Collaborative Clustering with back ground knowledge (Forestier et al.), Comparison between two Hierarchical Clusterings (Fowlkes and Mallows), Data Clustering: k-means (Jain) , Clustering algorithms in biomedical research (Xu and Wunsch) etc. Spectral clustering and the high-dimensional stochastic block model was addressed by Rohe et al. (2011). Nguyen et al. (2012) worked on Clustering with Multi-viewpoint-Based Similarity Measure and Fan and Albert (2013) focused on mining Big Data. Ghosh and Dubey (2013) worked on Comparative Analysis of K-Means and Fuzzy C Means Algorithms. Chen et al. (2014) Improved graph clustering. A Review on Big Data Clustering was addressed by Shirkhorshidi et al. (2014). Wu et al. (2014) concentrated on data mining with big data.

The Ministry of Tourism of the Government of India in 2009-10, launched a special initiative called “Hunar Se Rozgar Tak” (HSRT), for creation of employable skills amongst 8th pass youths belonging to economically weaker strata of the Indian society. The programme is fully funded by the Ministry of Tourism, India. Ministry of Tourism, Government of India published a progress report of HSRT in 2016. The report consisted of several tables displaying the number of trainees who opted for each of four separate courses namely, Food Processing (FP), Food and Beverage (F&B), Bakery and Processing (B&P) and House Keeping Unit (HKU) in each of the different institutes situated in different parts of India. These tables containing the above mentioned data generated a curiosity within the authors of this paper to analyse, so that the underlying facts relating to the training patterns, training institutes, time effects etc may be revealed. As described in the subsequent sections, clustering techniques and some significance testing have been employed. To the best of our knowledge and information, this kind of application of clustering techniques to group several training institutes and subsequent analysis has not been done before.

Objectives:-

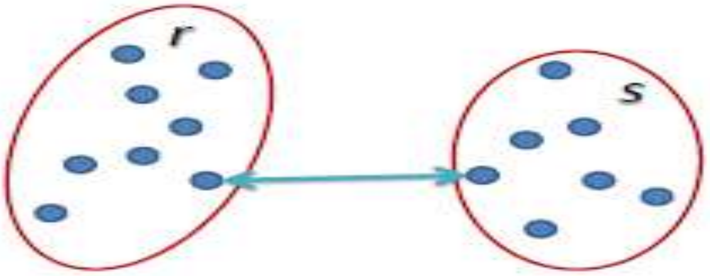
The main objectives of this work are as follows:

1. To study the similarity or differences between the training patterns provided by the different institutes under HSRT.

2. To judge whether time has played any role in changing the pattern of allocation of trainees to different courses of these institutes.
3. To find whether Geographical position of an institute plays any role in determining the relative frequencies of trainees opting for different types of training.
4. To determine whether any significant changes in the training infrastructure / condition of the different training institute under HSRT occur with the passage of time.

Methodology:-

Secondary data showing the conditions of different institutes (under HSRT) in terms of the number of trainees doing four different courses (FP, F&B, B&P and HKU) has been collected from progress report for HSRT published by Ministry of Tourism, Government of India in 2016. With this above mentioned data an effort has been made to place the institutes into different clusters, based on the similarity of those institutes. We have used Agglomerative Hierarchical Clustering with single linkage. Agglomerative hierarchical methods start with individual objects. Thus there are initially as many clusters as objects. The most similar objects are first grouped and these initial groups are merged according to their similarities. Eventually as the similarity decreases all sub groups are fused into a single cluster. In single linkage clustering the link between two clusters is made by a single pair of elements, namely those two elements (one in each cluster) that are closest to each other.



$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

Fig 1:- Single linkage Clustering.

On doing Hierarchical Agglomerative Clustering with the help of SPSS 16.0, dendrograms have been obtained. The dendrograms reveal a rough idea of the possible number of clusters (say N) in which the different institutes can be placed. To determine the exact nature of membership in those N clusters, K-mean clustering has been done. MacQueen suggested the term K-mean for describing one of his algorithm that assigns each item to the cluster having the nearest centroids (means). In our work we have taken the value of K equal to N. These N clusters have been ordered on the basis of the numbers of members in each of them and they are named as C_1, C_2, ..… , C_N, following the descending order of number of members in each of them. The same procedure is repeated for each of the four years from 2013 to 2016. Then to study whether the compositions of the clusters have undergone significant changes with the passage of time, comparisons have been made between the clustering compositions of the different years. To test whether the change (if any) is significant the help of chi square test has been employed.

Analysis and Findings

As already mentioned in the Introduction section, our dataset consist of data regarding the frequencies of trainees opting for each of the four courses for each of the institutes under HSRT. Initially, with this available data, Hierarchical Agglomerative Clustering of the institutes has been done with the help of SPSS 16.0. Four different dendrograms corresponding to four different years have been obtained. Due to shortage of space, only one of them is displayed below.

Dendrogram using Average Linkage (Between Groups)

Rescaled Distance Cluster Combine

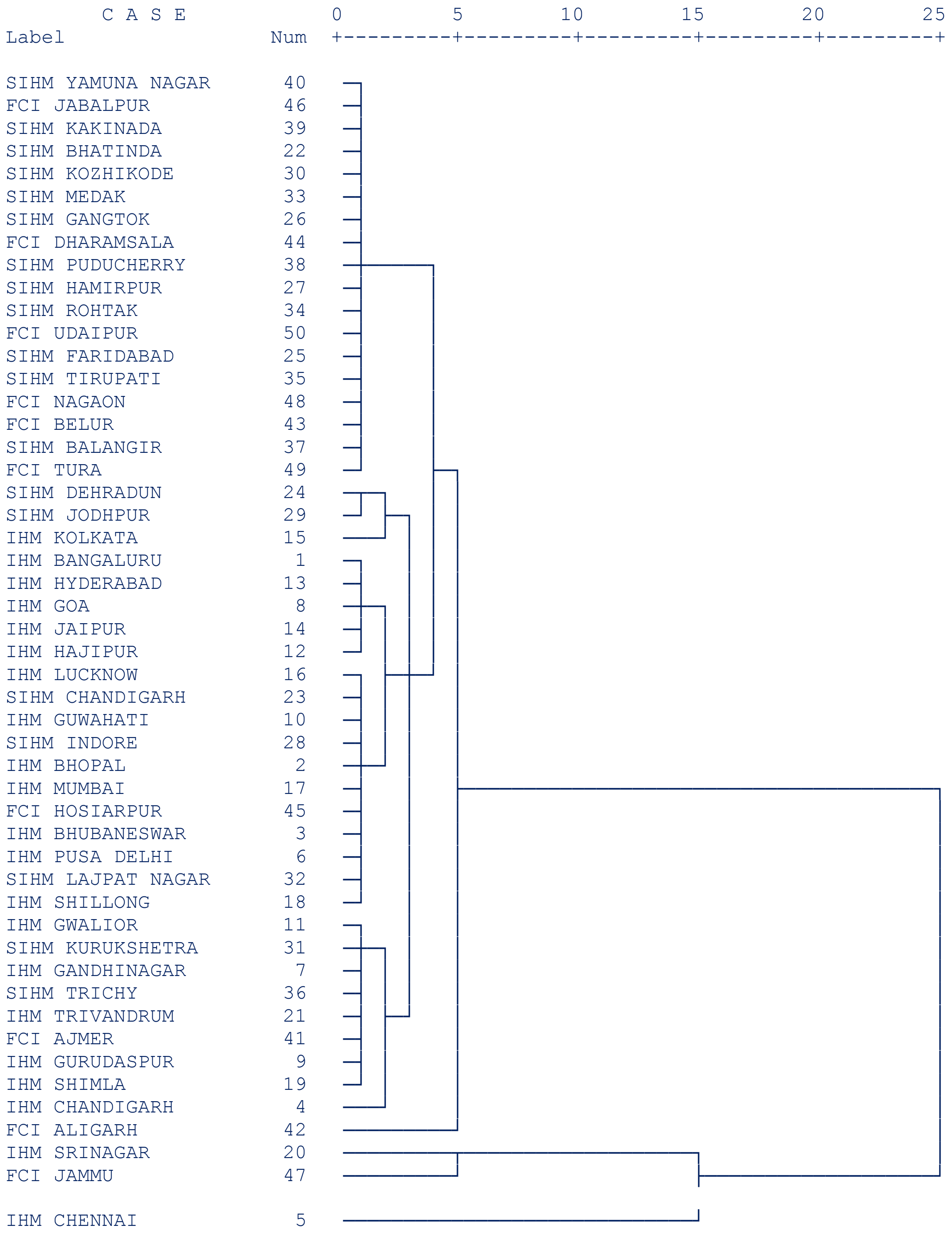


Fig 1:- Dendrogham.

From the dendrogram it is evident that if we permit the maximum distance between any two elements of the same cluster to be a reasonably considerable distance (say 4 units), then approximately 7 different clusters are noticed. To determine the exact nature of membership in those 7 clusters, K-mean clustering (taking K = 7) has been done, using SPSS 16.0. These 7 clusters have been ordered according to the number of members in each cluster (i.e., the cluster having maximum number of members is named C_1 and the cluster having minimum number of members is named C_7). The same procedure has been followed for each year.

The seven clusters obtained in each of the four years are presented [using the format: Cluster name/number (name of the members)(number of members in the cluster)] below.

Clusters in 2013

C_1 (SIHM BHATINDA,SIHM CHANDIGARH, SIHM FARIDABAD, SIHM GANGTOK, SIHM HAMIRPUR, SIHM KOZHIKODE, SIHM ROTHAK, SIHM PUDUCHERRY, SIHM KAKINADA, SIHM YAMUNANAGAR, FCI BELUR, FCI DHARAMSALA, FCI JABALPUR, FCI NAGAON, FCI TURA) (15)

C_2 (IHM CHANDIGARH, IHM GANDHINAGAR, IHM GURUDASPUR, IHM GUWAHATI, IHM GWALIOR, IHM LUCKNOW, IHM MUMBAI, IHM SHIMLA, IHM TRIVANDRUM, SIHM KURUKSHETRA, FCI AJMER) (11)

C_3 (IHM GOA, IHM HAJIPUR, IHM JAIPUR, IHM KOLKATA, SIHM INDORE, SIHM JODHPUR, SIHM LAJPAT NAGAR, SIHM MEDAK, FCI UDAIPUR) (9)

C_4 (IHM BANGALURU, IHM BHUBANESWAR, IHM CHENNAI, IHM HYDERABAD, SIHM DEHRADUN, SIHM TIRUPATI, SIHM TRICHY, SIHM BALANGIR) (8)

C_5 (IHM BHOPAL, IHM PUSA DELHI, IHM SHILLONG, FCI ALIGARH, FCI HOSIARPUR) (5)

C_6 (IHM SRINAGAR) (1)

C_7 (FCI JAMMU) (1)

Final Cluster Centers							
Courses ↓	Cluster						
	C_1	**C_2**	**C_3**	**C_4**	**C_5**	**C_6**	**C_7**
FP	221.67	883.00	80.87	117.88	457.00	183.20	241.82
FB	181.67	707.00	43.20	127.62	439.00	62.20	45.18
BP	.00	124.00	11.33	30.00	.00	146.60	18.18
HKU	11.67	113.00	4.73	23.62	105.00	42.80	6.36

Table 1:- Cluster Centers – 2013.

Table 1 displays the average frequencies of trainees in each of the clusters corresponding to each of the courses in 2013

Clusters in 2014

C_1 (SIHM BHATINDA, , SIHM FARIDABAD, SIHM GANGTOK, SIHM KOZHIKODE, SIHM MEDAK, SIHM TIRUPATI, SIHM BALANGIR, SIHM PUDUCHERRY, SIHM KAKINADA, SIHM YAMUNANAGAR, FCI BELUR, FCI DHARAMSALA, FCI JABALPUR, FCI TURA, FCI UDAIPUR) (15)

C_2 (IHM CHANDIGARH, IHM GANDHINAGAR, IHM GWALIOR, IHM GURUDASPUR, IHM GUWAHATI, IHM SHIMLA, SIHM INDORE, SIHM JODHPUR, IHM TRIVANDRUM, SIHM KURUKSHETRA, FCI AJMER, SIHM TRICHY) (12)

C_3 (IHM BHOPAL, IHM BHUBANESWAR, IHM PUSA DELHI, IHM GOA, IHM LUCKNOW, IHM MUMBAI, IHM SHILLONG, SIHM LAJPAT NAGAR, FCI ALIGARH, FCI HOSIARPUR, SIHM CHANDIGARH) (11)

C_4 (IHM BANGALURU, IHM HAJIPUR , IHM HYDERABAD, IHM JAIPUR, IHM KOLKATA, SIHM DEHRADUN, SIHM HAMIRPUR, SIHM ROTHAK, FCI NAGAON) (9)

C_5 (IHM CHENNAI) (1)

C_6 (IHM SRINAGAR) (1)

C_7 (FCI JAMMU) (1)

Final Cluster Centers							
Courses ↓	Cluster						
	C_1	**C_2**	**C_3**	**C_4**	**C_5**	**C_6**	**C_7**

FP	602.00	73.20	239.09	316.83	730.00	164.11	809.00
FB	561.00	40.73	57.82	31.58	19.00	120.67	409.00
BP	103.00	5.00	149.18	20.00	148.00	62.67	47.00
HKU	7.00	18.33	23.64	23.83	.00	21.11	30.00

Table 2:- Cluster Centers – 2014.

Table 2 displays the average frequencies of trainees in each of the clusters corresponding to each of the courses in 2014

Clusters in 2015

C_1 (IHM JAIPUR, IHM TRIVANDRUM, IHM SRINAGAR, SIHM BHATINDA, SIHM CHANDIGARH, SIHM FARIDABAD, SIHM GANGTOK, SIHM HAMIRPUR, SIHM INDORE, SIHM KOZIKODE, SIHM MEDAK, SIHM BALANGIR, SIHM PUDUCHERRY, SIHM KAKINADA, SIHM YAMUNA NAGAR, FCI BELUR, FCI DHARAMSALA, FCI UDAIPUR, SIHM TIRUPATI) (19)

C_2 (IHM GANDHINAGAR, IHM GURUDASPUR, IHM GUWAHATI, IHM KOLKATA, IHM SHIMLA, SIHM DEHRADUN, SIHM JODHPUR, SIHM ROTHAK, SIHM TRICHY, FCI JABALPUR) (10)

C_3 (IHM BHOPAL, IHM BHUBENSWAR, IHM PUSA DELHI, IHM GOA, IHM LUCKNOW, IHM MUMBAI, IHM SHILLONG, SIHM LAJPAT NAGAR, FCI HOSIARPUR) (9)

C_4 (IHM BANFALURU, IHM HAJIPUR, IHM HYDERABAD, FCI NAGAON, FCI TURA) (5)

C_5 (IHM CHANDIGARH, IHM CHENNAI, IHM GWALIOR, SIHM KURUKSHETRA, FCI AJMER) (5)

C_6 (FCI ALIGARH) (1)

C_7 (FCI JAMMU) (1)

Final Cluster Centers							
Courses ↓	Cluster						
	C_1	**C_2**	**C_3**	**C_4**	**C_5**	**C_6**	**C_7**
FP	306.00	232.67	134.60	80.95	482.80	801.00	301.10
FB	91.00	29.22	127.40	28.16	48.00	754.00	99.20
BP	179.00	167.67	101.80	7.74	26.60	69.00	8.60
HKU	286.00	8.00	76.40	7.05	23.80	.00	15.40

Table 3:- Cluster Centers – 2015.

Table 3 displays the average frequencies of trainees in each of the clusters corresponding to each of the courses in 2015

Clusters in 2016

C_1 (IHM JAIPUR, IHM SHIMLA, IHM TRIVANDURM, SIHM BHATINDA, SIHM CHANDIGARH, SIHM FARIDABAD, SIHM GANGTOK, SIHM HAMIRPUR, SIHM INDORE, SIHM KOZIKODE, SIHM LAJPAT NAGAR, SIHM MEDAK, SIHM ROHTAK, SIHM TIRUPATI, SIHM TRICHI, SHIM BALANGIR, SIHM KAKINADA, SIHM PUDUCHERRY, SIHM YAMUNANAGAR, FCI BELUR, FCI DHARAMSALA, FCI NAGAON, FCI TURA, FCI UDAIPUR) (24)

C_2 (IHM BHUBENESWAR, IHM GOA, IHM GUWAHATI, IHM LUCKNOW, IHM MUMBAI, IHM SHILLONG, FCI ALIGARH, FCI HOSSIARPUR) (8)

C_3 (IHM BHOPAL, IHM BANGALURU, IHM HYDERABAD, IHM HAJIPUR, IHM KOLKATA, IHM SRINAGAR) (6)

C_4 (IHM GANDHINAGAR, IHM GURUDASPUR, SIHM DEHRADUN, SIHM JODHPUR, FCI JABALPUR) (5)

C_5 (IHM PUSA DELHI, IHM GUWALIOR, SIHM KURUKSHETRA) (3)

C_6 (IHM CHANDIGARH, IHM CHENNAI, FCI AJMER) (3)

C_7 (FCI JAMMU) (1)

Final Cluster Centers	
	Cluster

Courses ↓	C_1	C_2	C_3	C_4	C_5	C_6	C_7
FP	404.20	179.50	108.96	606.00	251.75	530.33	502.00
FB	199.00	207.17	30.08	.00	31.75	62.33	745.00
BP	4.80	74.17	19.67	11.67	178.00	159.00	57.00
HKU	.00	96.00	16.33	10.67	24.25	.00	212.00

Table 4:- Cluster Centers – 2016.

Table 4 displays the average frequencies of trainees in each of the clusters corresponding to each of the courses in 2016

From the clusters of each year visible above, it is noticed that most of the members remained in the same cluster with more or less same set of co-members over the years, which implies that either all these member institutes have had the same conditions including infrastructure all throughout this time period (2013-16) or else their conditions have jointly changed in the same manner, i.e., there has not been any special noticeable change for any of the specific individual institute so that it can be isolated from others. Of course at a glance, some characteristic features are observed like

1. FCI JAMMU has its own typical characterised identity which is completely different from any of the other institutes in consideration. This is demonstrated by the fact that it has always remained as a unique member of a single cluster in all the four years. One may further study the infrastructure and conditions of this institute to analyze its uniqueness.
2. The same effect as discussed for FCI JAMMU is visible to a marginally leser extent in case of IHM SRINAGAR. It appears in a single cluster in year 2013 and 2014.
3. Apart from the above two points the data reveals that there are a few institutes which have always been a part of a small clusters. These institutes are jointly different in their characteristics from the majority of other institutes (examples: IHM CHANDIGARH, IHM PUSA DELHI, IHM CHENNAI etc).

To conclusively judge whether the clustering of these institutes has significantly changed over the years, comparison between the clustering pattern of one year with another year, taking different pairs of years have been done using χ^2 test.

Comparison between 2013 and 2014

One may be interested to study whether the clustering of the institutes in 2013, in terms of the composition of the cluster has undergone a significant change in 2014. If there is no significant change then it would imply that with the passage of time, the member institutes have had the same conditions including infrastructure all throughout this time period 2013-14 or else their conditions have jointly changed in the same manner, i.e., there has not been any special noticeable change for any of the individual institute.

H_0: Clustering of 2014 is independent of clustering of 2013.
H_1: Clustering of 2014 is dependent of clustering of 2013.
Number of Cluster members in 2014

Number of Cluster members in 2013	CLUSTER	C_1	C_2	C_3	C_4	C_5	C_6	C_7	Total numbers of members
	C_1	11	0	1	3	0	0	0	15
	C_2	0	9	2	0	0	0	0	11
	C_3	2	2	2	3	0	0	0	9
	C_4	2	1	1	3	1	0	0	8
	C_5	0	0	5	0	0	0	0	5
	C_6	0	0	0	0	0	1	0	1
	C_7	0	0	0	0	0	0	1	1
	Total numbers of members	15	12	11	9	1	1	1	50

Table 5:- Transition frequency – 2013-2014.

The Table 5 shows the transition or no transition of members from cluster of 2013 to cluster of 2014. For e.g., the figure 2 in the cell (2, 3) indicates two members of cluster C_2 of 2013 has shifted to cluster C_3 of 2014.
Here, degrees of freedom, df = (7-1)(7-1)=36

Calculated χ^2 = 161.75
For 36 degrees of freedom at 5% level of significance, tabulated χ^2 = 50.998 i.e. cal χ^2 >> tab χ^2.
So, H_0 is rejected. So there is a dependency between the cluster of 2013 and 2014. So passage of time does not have any significant impact on clustering. The relative nature of the training provided by the different institutes did not undergo any significant alteration in 2014 as compared to 2013.

Comparison between 2014 and 2015
H_0: Clustering of 2015 is independent of clustering of 2014.
H_1: Clustering of 2015 is dependent of clustering of 2014.
Number of Cluster members in 2015

	CLUSTER	C_1	C_2	C_3	C_4	C_5	C_6	C_7	Total numbers of members
Number of Cluster members in 2014	C_1	13	1	0	1	0	0	0	15
	C_2	2	6	0	0	4	0	0	12
	C_3	1	0	9	0	0	1	0	11
	C_4	2	3	0	4	0	0	0	9
	C_5	0	0	0	0	1	0	0	1
	C_6	1	0	0	0	0	0	0	1
	C_7	0	0	0	0	0	0	1	1
	Total numbers of members	19	10	9	5	5	1	1	50

Table 6:- Transition frequency – 2014-2015.

The Table 6 shows the transition or no transition of member from cluster of 2014 to cluster of 2015. For e.g., the figure 13 in the cell (1, 1) indicates 13 members of cluster C_1 of 2014 has remained to cluster C_1 of 2015, but 1 member has shifted to cluster C_2 (cell (1,2)) and another 1 member has shifted to cluster C_4 (cell (1,4)) of 2015.

Here, degrees of freedom, df = (7-1)(7-1)=36
Calculated χ^2 = 142.132

For 36 degrees of freedom at 5% level of significance, tabulated χ^2 = 50.998 i.e. cal χ^2 >> tab χ^2.
So, H_0 is rejected. So there is a dependency between the cluster of 2014 and 2015. So passage of time does not have any significant impact on clustering. The relative nature of the training provided by the different institutes did not undergo any significant alteration in 2015 as compared to 2014.

Comparison between 2015 and 2016
H_0: Clustering of 2016 is independent of clustering of 2015.
H_1: Clustering of 2016 is dependent of clustering of 2015.
Number of Cluster members in 2016

	CLUSTER	C_1	C_2	C_3	C_4	C_5	C_6	C_7	Total numbers of members
Number of Cluster members in 2015	C_1	18	0	1	0	0	0	0	19
	C_2	3	1	1	5	0	0	0	10
	C_3	1	6	1	0	1	0	0	9
	C_4	2	0	3	0	0	0	0	5
	C_5	0	0	0	0	2	3	0	5
	C_6	0	1	0	0	0	0	0	1
	C_7	0	0	0	0	0	0	0	1
	Total numbers of members	24	8	6	5	3	3	1	50

Table 7:- Transition frequency – 2015-2016.

The Table 7 shows the transition or no transition of member from cluster of 2015 to cluster of 2016. With this data also χ^2- test has been done and the following result is found .
Calculated χ^2 = 158.557

For 36 degrees of freedom at 5% level of significance, tabulated χ^2 = 50.998 i.e. cal χ^2 >> tab χ^2.

So, H_0 is rejected. So there is a dependency between the cluster of 2015 and 2016. So passage of time does not have any significant impact on clustering. The relative nature of the training provided by the different institutes did not undergo any significant alteration in 2016 as compared to 2015.

Comparison between 2013 and 2016
H_0: Clustering of 2016 is independent of clustering of 2013.
H_1: Clustering of 2016 is dependent of clustering of 2013.

Number of Cluster members in 2016

Number of Cluster members in 2013	CLUSTER	C_1	C_2	C_3	C_4	C_5	C_6	C_7	Total numbers of members
	C_1	14	0	0	1	0	0	0	15
	C_2	2	3	0	2	2	2	0	11
	C_3	5	1	2	1	0	0	0	9
	C_4	3	1	2	1	0	1	0	8
	C_5	0	3	1	0	1	0	0	5
	C_6	0	0	1	0	0	0	0	1
	C_7	0	0	0	0	0	0	1	1
	Total numbers of members	24	8	6	5	3	3	1	50

Table 8:- Transition frequency – 2013-2016.

The Table 8 shows the transition or no transition of member from cluster of 2015 to cluster of 2016. With this data also χ^2- test has been done and the following result is found .
Calculated $\chi^2 = 96.022$

For 36 degrees of freedom at 5% level of significance, tabulated $\chi^2 = 50.998$ i.e. cal $\chi^2 >>$ tab χ^2.
So, H_0 is rejected. So there is a dependency between the cluster of 2013 and 2016. So passage of time does not have any significant impact on clustering in long term also. The relative nature of the training provided by the different institutes did not undergo any significant alteration in 2016 as compared to 2013.

Thus no particular institute can be singled out for showing any contrasting change in its training pattern offered over the years

Conclusion:-

This paper is the outcome of a simple effort to apply clustering techniques in comparing the training patterns (as well as the pattern of allocating different trainees to different courses) provided by the different training institutes under a Government project. To the best of our knowledge/information, this kind of study involving clustering of training institutes has not been done before. Moreover, the data set available (although fully authentic) had limitations due to the fact that it lacked variety and only data regarding frequencies of trainees opting for each of four different courses in each of the several institutes were available. However, analysis using clustering techniques and Chi-square test has revealed some interesting results. The study reveals that the training pattern has not undergone any significant change during the period 2013-2016. In Government training institutes the infrastructural condition does not change much with the passage of time. Some training institutes like FCI JAMMU and IHM SRINAGAR have unique characteristics. One may undergo further study on the infrastructure and other conditions of these institutes to analyze their uniqueness. In this regard it is also felt that the geographical positions of these two institutes have a crucial role to play behind this uniqueness.

Reference:-

1. Agrawal, R., Johannes, G., Dimitrios, G., & Raghavan, P. (1998). Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. SIGMOD Conference, 94-105.
2. Brendan, J. F., & Dueck, D. (2007). Clustering by passing messages between data points. Science, 315, 972–976.

3. Castro, V. E., & Yang, J. (2000). A Fast and robust general purpose clustering algorithm. International Conference on Artificial Intelligence.
4. Celeux, G., & Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. Computational statistics & Data analysis, 14(3), 315-332.
5. Chen, Y., Sanghavi, S., & Xu, H. (2014). Improved graph clustering. IEEE Transactions on Information Theory, 60(10), 6440-6455.
6. Dolnicar, S. (2003). Using Cluster Analysis for Market Segmentation–Typical Misconceptions, Established Methodological Weaknesses and Some Recommendations for Improvement. Journal of Marketing Research, 11(2), 5-12.
7. Dunn, J. C.(1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics, 3(3), 32-57.
8. Faceili, K., Carvalho, A. D., & Souto, D. (2006). Multiobjective Clustering ensemble. International Conference on Hybrid Intelligent Systems.
9. Fan, W., & Albert, B. (2013). Mining Big Data: Current Status and Forecast to the Future. ACM SIGKDD Explorations Newsletter, 14(2), 1-5.
10. Forestier, G., Gancarski, P., & Wemmert, C. (2010). Collaborative Clustering with back ground knowledge. Data and Knowledge Engineering, 69(2), 211–228.
11. Fowlkes, E. B., & Mallows, C. L. (2010). A Method for Comparing Two Hierarchical Clusterings. Journal of the American Statistical Association, 78(383), 553–569.
12. Fraley, C., & Raftery, A. E. (1998). How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. Technical Report No. 329, Department of Statistics University of Washington.
13. Gath, I., & Geva, A. (1989). Unsupervised optimal fuzzy clustering. IEEE Transaction on Pattern Analysis and Machine Intelligence, 11(7), 773–781.
14. Ghosh, S., & Dubey, S. K. (2013). Comparative Analysis of K-Means and Fuzzy C Means Algorithms. International Journal of Advanced Computer Science and Applications, 4(4), 35-39.
15. Ghuman, S. S. (2016). Clustering Techniques – A Review. International Journal of Computer Science and Mobile Computing, 5(5), 524-530.
16. Han, J., Kamber, M. & Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.
17. Handl, J., & Knowles, J. (2007). An evolutionary approach to Multiobjective clustering. IEEE Transaction on Evolutionary Computation, 11(1), 56-76.
18. Hanson, S., & Bauer, M. (1989). Conceptual clustering, categorization and polymorphy. Machine Learning Journal, 3(4), 343-372.
19. Jain, A. K. (2010). Data Clustering: 50 years beyond k-means. Pattern Recognition Letters, 31(8), 651–666.
20. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: A review. ACM Computing Surveys, 31(3), 264-323.
21. King, B. (1967). Step-wise Clustering Procedures. Journal of American Statistical Association , 69(317), 86-101.
22. Krishnapuram, R., & Keller, J. (1993). A possibilistic approach to clustering. IEEE Transaction on Fuzzy Systems, 1(2), 98–110.
23. Law, M. K., Topchy, A., & Jain, A. K. (2004). Multiobjective Data Clustering. IEEE Conference on Computer Vision and Pattern Recognition, 2, 424-430.
24. Luxburg, U. (2007). A tutorial on spectral clustering. Statistics and Computing, 17(4), 395-416.
25. MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. 5th Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1, 281-297.
26. Murtagh, F. (1984). A survey of recent advances in hierarchical clustering algorithms which use cluster centers. Computer Journal, 26(4), 354-359.
27. Nakayama, H., & Kagaku, N. (1998). Pattern classification by linear goal programming and its extensions. Journal of Global Optimization, 12(2), 111–126.
28. Nguyen, D. T., Chen, L., & Chan, C. K. (2012). Clustering with Multi-viewpoint-Based Similarity Measure. IEEE Transactions on Knowledge and Data Engineering, 24(6), 988-1001.
29. Rohe, K., Chatterjee, S., & Yu, B. (2011). Spectral clustering and the high-dimensional stochastic block model. The Annals of Statistics, 39(4), 1878-1915.
30. Rokach, L. (2005). Clustering Methods. Data Mining and Knowledge Discovery Handbook, 331-352.
31. Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M.J., Ding, W., Lin, C.T. (2017) A review of clustering techniques and developments. Neurocomputing,Accepted Manuscript

32. Sheikholeslami, G., Chatterjee, S., & Zhang, A. (2000). WaveCluster: a wavelet-based clustering approach for spatial data in very large databases. The International Journal on Very Large Data Bases, 8(3-4), 289-304.
33. Shirkhorshidi, A. S., Aghabozorgi, S., Wah, T. Y., & Herawan, T. (2014). Big Data Clustering: A Review. Lecture Notes in Computer Science, 8583, 707-720.
34. Sneath, P., & Sokal, R. (1973). Numerical Taxonomy. W.H. Freeman Co, San Francisco, CA.
35. Urquhart, R. (1982). Graph-theoretical clustering based on limited neighborhood sets. Pattern Recognition, 15(3), 173-187.
36. Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58(301), 236-244.
37. Wolpert, D. H., & Macready, W. G. (1997). No Free Lunch Theorem for Optimization. IEEE Transactions on Evolutionary Computation, 1(1), 67-82.
38. Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. IEEE Transaction on Knowledge and Data Engineering, 26(1), 97-107.
39. Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transaction on Neural Networks, 16(3), 645–678.
40. Xu, R., & Wunsch, D. (2010). Clustering algorithms in biomedical research: a review. IEEE Reviews in Biomedical Engineering, 3, 120–154.
41. Zahn, C. T. (1971). Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Transaction on Computer, C-20(1), 68-86.