Hierarchical clustering of students' software project development technique - Local Vs. Global Approach

Purnima Kumari Srivastava[#1], Mamata Pandey[#2], Vandana Bhattacharjee*[3]

[#] Department of IT, Ranchi University
Ranchi Women's College, Ranchi, India
[1]purnima.srivas@yahoo.com
[2] pandeymamata78@gmail.com
[*] Department of Computer Science and Engineering, BIT Mesra
Ranchi, India
[3]vbhattacharya@bitmesra.ac.in

***Abstract*—Clustering is an unsupervised machine learning process that creates clusters such that data points inside a cluster are close to each other and also apart from data points in other clusters. There are many clustering technique to group the data objects on the basis of similarity, distance and common neighbour. The hierarchical clustering technique is one of them. This paper describes the comparative result for evaluating the students' projects by local and global approach of hierarchical clustering technique. The clustering result was validated by a panel of domain experts.**

Keyword- Hierarchical Clustering, Local approach, Global approach

I. INTRODUCTION

Clustering is the technique of grouping data so that the data in each group share similar characteristics and patterns. There are many techniques which are used to form clusters. Most commonly, hierarchical and partitional techniques are used to group similar data objects in one cluster. The clusters are formed by the similarity measures [1].

Due to enhancement in computer technology, programming language, software and hardware tools the difficult task are easily solved. The people are more interested in finding new way of performing the task in simple and easiest manner. The employment level of every country is also increasing. The attendance of students in Universities and schools are gradually increasing. Everyday many projects belonging to various categories are being gradually developed through teaching institution. It is very difficult to diagnose the categories of projects developed by students traditionally (manually) on the basis of few attributes like title of project, Operating System, Programming languages etc.

This paper provides the framework for evaluation of students' projects of different institution by using hierarchical clustering techniques. This paper also provides the help in analyzing the students' projects. The projects which are similar in their project attribute can group together. The teacher can differentiate the projects. The students have similar ideas can also share their view. The groupings of projects are based on similarity or dissimilarity basis

The rest of paper is organized as follows. In Section II, the hierarchical clustering technique based on distance measure is explained. In Section III, the hierarchical clustering technique based on link (ROCK) is explained. In section IV, local and global approach of hierarchical clustering is explained. In Section V, the similarity measure (jaccard coefficient) is explained. In section VI the methodology for implementation of local and global hierarchical clustering on students' projects is explained. Section VII explains the experimental method for grouping students based on similarity measures. The conclusion is presented in Section VIII.

II. HIERARCHICAL CLUSTERING

Hierarchical clustering creates the hierarchical decomposition of database. The algorithm iteratively split the database into smaller subset, until some termination condition is satisfied [2]. The hierarchical clustering algorithms do not need *k* as an input parameter, which is an advantage over partitioning algorithms. The hierarchical decomposition can be represented by dendrogram in two ways [2].

I. Bottom-up (agglomerative) approach
II. Top-down (Divisive) approach.

The basic agglomerative, hierarchical clustering algorithm works as following ways [3]

Initially each object is placed in a unique cluster. For each pair of clusters, some value of dissimilarity or distance is computed. For instance, the distance may be in minimum distances (Single linkage) in the current clustering are merged, until the whole data sets forms a single cluster [3].

III. ROCK (ROBUST CLUSTERING WITH LINKS)

ROCK is an adaptation of an agglomerative hierarchical clustering algorithm. An agglomerative hierarchical clustering normally uses distance-based representatives (Euclidean distance) to determine the similarity between clusters. It is observed that such similarity function tend to merge clusters which have disjoint set. Moreover it is not possible to extend the concept of centriod of categorical attributes. ROCK makes use of links for defining similarity. The number of links between two tuples is the number of common neighbors they have in data set. Starting with each tuple in its own clusters, the two closest clusters are merged until the required numbers of clusters are obtained [4].

A. Clustering Paradigm

The following are the ROCK clustering paradigm.

1) *Neighbours.:* An object neighbours are those objects that are considerably similar to it. Given a threshold θ between 0 and 1, a pair of points O_i,O_j are defined to be neighbours if the following holds:

$Sim(O_i,O_j) \geq \theta$.

The value of sim is [0,1], with larger values indicating that the points are more similar.

2) *Links:* The link (O_i,O_j) between the objects is defined as the number of common neighbors between O_i and O_j. If the link (O_i,O_j) is large, then it is more probable that O_i and O_j belong the same cluster.

3) *Goodness Measure:* ROCK uses link-based agglomerative hierarchical clustering approach. It starts with a singleton objects as an individual class and merge progressively merge the clusters based on goodness criteria, determined by the link structure [5]. Finally, the clusters involving only the sample objects are used to assign the remaining data objects on the disk to appropriate clusters.

For a pair of clusters C_i,C_j, let link[C_i,C_j] store the number of cross links between clusters Ci and C_j, that is, $O_{pq} \in C_i, O_r \in C_j$ link(pq,pr). Then, the goodness measure g(C_i,C_j) for merging clusters C_i,C_j is as follows[4].

$g(C_i,Cj) = \frac{link[Ci,Cj]}{(ni+nj)^{1+2f(\theta)} - ni^{1+2f(\theta)} - nj^{1+2f(\theta)}}$, ni and nj are the number of points in each cluster.

Where, $f(\theta) = \frac{1-\theta}{1+\theta}$, $\theta < 1$

The pair of clusters for which the goodness measure is maximum is the best pair of clusters to be merged at any given step.

IV. LOCAL AND GLOBAL APPROACH OF HIERARCHICAL CLUSTERING

The data objects are clusters by the distance, similarity or linked basis. We can use hierarchical clustering approach to cluster the data objects based on distance, similarity or common links. When we use distance or similarity for clustering the data objects, it is called local approach and if the clusters are formed by measuring the common links among data objects it is called global approach [4].

The similarity measure between a pair of points takes account characteristics of the points themselves; it is local approach of clustering. The link based approach captures the global knowledge of the neighboring data points into the relationship between individual pair of points [6].

V. SIMILARITY MEASURES

A. Jaccard coefficient

Clustering algorithms usually a distance metric based (e.g. Euclidean) similarity measure which involves grouping data into classes or clusters so those objects within the same cluster are similar whereas objects in different clusters are relatively dissimilar. But this metric is not suitable for categorical data. The similarity among categorical data can be measured by jaccard coefficient[5].

$\text{similarity } (O_i,O_j) = \frac{\text{Number of attributes in common}}{\text{Number of attributes in both}} = (O_i \cap O_j) / (O_i \cup O_j)$

The more attributes that the two transactions O_1 and O_2 have in common, that is, the larger

$| O_1 \cap O_2 |$ is, the more similar they are. Dividing by $| O_1 \cup O_2 |$ is the scaling factor which ensures that θ (threshold distance) is between 0 and 1. Thus , the above equation computes the relative closeness based on Objects attributes appearing in both data objects O_1 and O_2.The similarity among data objects (O_i,Oj) is [0,1]. Higher the value means higher similarity.

VI. METHODOLOGY

The following steps are taken to compare the local and global approach of hierarchical clustering technique as shown in Fig. 1. The comparison is done on the basis of execution time and space taken by both approach.

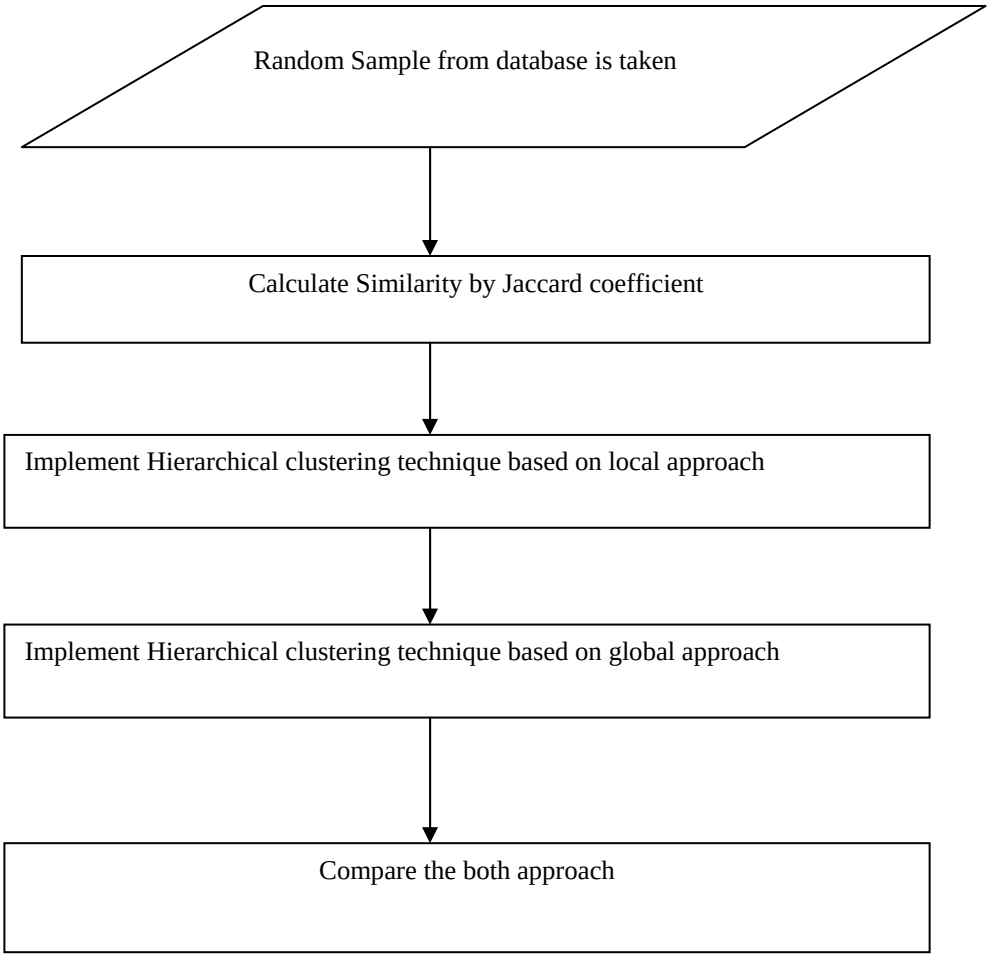

Fig. 1. Process Architecture

VII. ILLUSTRATION BY EXAMPLE

To compare the local and global approach of Hierarchical clustering ten sets of raw data of students are taken for observation. The following set of nine attributes has been considered in the clustering process[6]. These attributes are taken for preparing final year projects in computer science in various institutions..

→Objective(1)
→Feasibility Study(2)
→Project Scheduling(3)
→Software Requirement specification (4)
→Data Flow Diagram(5)
→Database Diagram(6)
→Coding(7)
→Testing(8)
→Report Generation (9).

The symbol 1, 2, 3…. 9 are used for indicating the above nine attributes. Groups of ten students are selected as example. The presences of attributes in the students' projects are defined by the numeric symbols. For example if "Student6" mentioned feasibility report and database diagram in his project(S6) then the symbol 2 and 6 will be used to represent the presence of attributes Feasibility report and Project Scheduling. The symbols S1, S2,S3….. S10 are used to represent students' projects. The projects are similar if they have at least one common attributes.

S1={ 1, 2, 3, 5, 6}
S2={ 1, 3, 5, 8, 9}
S3={ 2, 3, 4, 5, 6,8}
S4={ 1, 3, 5}
S5={ 1, 5, 6}
S6={ 2, 3}
S7={ 7}
S8={ 3, 7}
S9={ 1,2}
S10={2,5,6}

A. Finding Similarity matrix by Jaccard coefficient

Based on above attributes, a Similarity matrix as shown as in Table I is created by using Jaccard coefficient.

TABLE I. SIMILARITY MATRIX of SIZE 10 X 10

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1	1	0.43	0.57	0.60	0.60	0.40	0	0.17	0.40	0.60
S2		1	0.38	0.60	0.33	0.17	0	0.17	0.17	0.14
S3			1	0.29	0.29	0.33	0	0.14	0.14	0.50
S4				1	0.50	0.25	0	0.25	0.25	0.20
S5					1	0	0	0	0.25	0.50
S6						1	0	0.33	0.33	0.25
S7							1	0.50	0	0
S8								1	0	0
S9									1	0.25
S10										1

B. Implementation of hierarchical clustering through local approach

The Agglomerative Hierarchical clustering technique based on single linkage similarity measure is used to find the similarity and dissimilarity among students' projects. Dissimilarity matrix (distance) is obtained by subtracting all value of possible pair of similarity matrix by 1 as shown in Table II. All projects initially belong to their own cluster. Clusters have minimum dissimilarity, are merged until the whole data sets forms a single cluster. In first iteration the distance between students' project S1 and S4 was minimum i.e. 0.40, so the student project S1 and S4 merged together in a cluster. We assigned new cluster name "S14". The size of dissimilarity matrix is reduces to 9 x 9 as shown in Table III. The name of cluster is assigned to demonstrate the result in simple way.

TABLE II. DISSIMILARITY MATRIX of SIZE 10 x 10

	S1	S2	S3	**S4**	S5	S6	S7	S8	S9	S10
S1	0	0.57	0.43	**0.40**	0.40	0.60	1	0.83	0.60	0.40
S2		0	0.62	0.40	0.67	0.83	1	0.83	0.83	0.86
S3			0	0.71	0.71	0.67	1	0.86	0.86	0.50
S4				0	0.50	0.75	1	0.75	0.75	0.80
S5					0	1	1	1	0.75	0.50
S6						0	1	0.67	0.67	0.75
S7							0	0.50	1	1
S8								0	1	1
S9									0	0.75
S10										0

In second Iteration again the minimum distance is 0.40 between S14 and S2 and hence students' project S14 and S2 merged together in a cluster. This process will be continue, until the whole data sets forms a single cluster.

Finally the cluster S78 is merged with cluster S142510369 at the distance 0.67. The result of entire process of clustering of students' project is represented by following dendrogram as in Fig. 2.

C. Implementation of hierarchical clustering through global approach

The global approach is based on links. We have used ROCK algorithm, which is the extension of agglomerative hierarchical clustering approach to find the similarity among students' projects.

We are selected θ=0.30(threshold distance). We applied this value to the similarity matrix as shown as Table I . The following Adjacency matrix will be generated when θ=0.30 as shown in Table IV.

TABLE III. DISSIMILARITY MATRIX of SIZE 9 x 9

	S14	**S2**	S3	S5	S6	S7	S8	S9	S10
S14	0	**0.40**	0.43	0.40	0.60	1	0.75	0.60	0.40
S2		0	0.62	0.67	0.83	1	0.83	0.83	0.86
S3			0	0.71	0.67	1	0.86	0.86	0.50
S5				0	1	1	1	0.75	0.50
S6					0	1	0.67	0.67	0.75
S7						0	0.50	1	1
S8							0	1	1
S9								0	0.75
S10									0

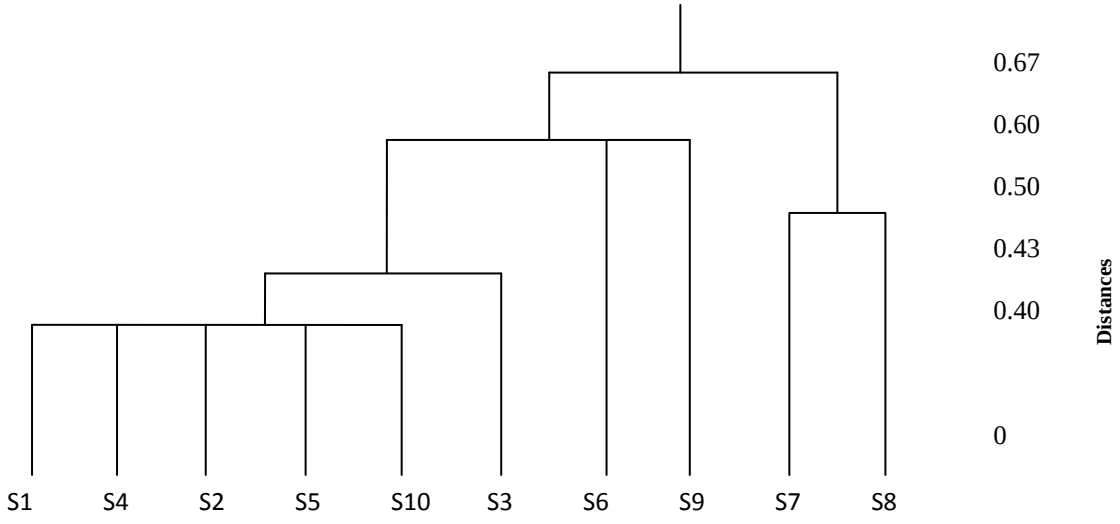

Fig . 2 Dendrogram of student projects based on similarity (local approach)

TABLE IV. ADJACENCY MATRIX of SIZE 10 x 10

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1	1	1	1	1	1	1	0	0	1	1
S2		1	1	1	1	0	0	0	0	0
S3			1	0	0	1	0	0	0	1
S4				1	1	0	0	0	0	0
S5					1	0	0	0	0	1
S6						1	0	1	1	0
S7							1	1	0	0
S8								1	0	0
S9									1	0
S10										1

By multiplying the adjacency table itself, A link matrix will be derived which shows the number of links (or common neighbors)as in Table V.

The goodness measure g(Ci,Cj) for merging clusters C_i,C_j is as follows.

$$g(C_i,C_j) = \frac{link[Ci,Cj]}{(ni+nj)^{1+2f(\theta)} - ni^{1+2f(\theta)} - nj^{1+2f(\theta)}}$$

Where, $f(\theta) = \frac{1-\theta}{1+\theta}$, $\theta < 1$

The goodness measure is calculated by above formula. The result of goodness measure is as follows as in Table VI.

TABLE V. LINK MATRIX A x A of SIZE 10 x 10

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1	-	5	5	4	5	4	0	1	3	4
S2		-	3	4	4	2	0	0	1	3
S3			-	2	3	3	0	1	2	3
S4				-	4	1	0	0	1	2
S5					-	1	0	0	1	3
S6						-	1	2	3	2
S7							-	2	0	0
S8								-	1	0
S9									-	1
S10										-

TABLE VI. GOODNESS MEASURE of LINK MATRIX of SIZE 10 x 10

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1	-	**2.253**	2.253	1.802	2.253	1.802	0	0.450	1.351	1.802
S2		-	1.351	1.802	1.802	0.901	0	0	0.450	1.351
S3			-	0.901	1.351	1.351	0	0.450	0.901	1.351
S4				-	1.802	0.450	0	0	0.450	0.901
S5					-	0.450	0	0	0.450	1.351
S6						-	0.450	0.901	1.351	0.901
S7							-	0.901	0	0
S8								-	0.450	0
S9									-	0.450
S10										-

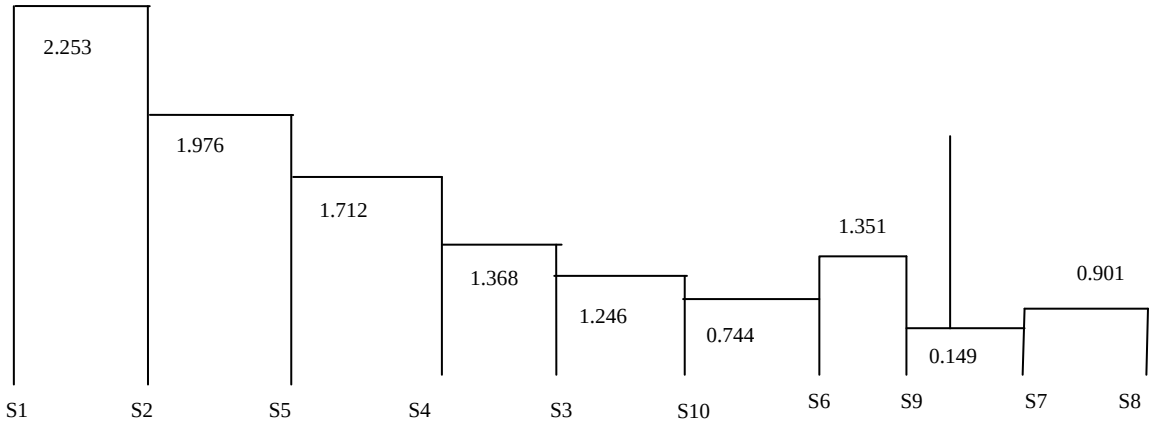

Fig. 3 Representation of merging of clusters based on goodness measure by global approach of hierarchical clustering technique

Clusters for which the goodness measure is maximum, is the best pair of clusters to be merged at any given step. The students' projects S1 and S2 has highest value of goodness measure so the merged in first iteration. The link matrix and goodness measure is again calculated. This process continues until we will find desired number of clusters or when there is a single cluster will be remained. The entire process of clustering through global approach of hierarchical technique is represented by Fig. 3.

D. Results of comparison with parameters

Execution time analysis for both approaches is done on the basis of the number of records that are considered for clustering and how much time is taken by this whole process as given in table Table VII. The analysis is also done on the number of iteration taken by each approach which is given in Table VIII

VIII. CONCLUSION

By a panel of experts it is found that the both local and global gives almost same result if there is a small data set(less than or equal to 10) having no outlier and where similarity will be measured through jaccard coefficient.

TABLE VII. COMPARISION with PARAMETER EXECUTION TIME

Number of record	Execution time for Hierarchical technique by local approach(ms.)	Execution time for Hierarchical technique by global approach(ms)
5	7	9
10	19	35

TABLE VIII. COMPARISION with PARAMETER – NUMBER OF ITERATION

Number of Record	Number of iteration by local approach	Number of iteration by global approach
10	330	550

The project S7's attribute is not found in other project except project S8.So the projects S7 is similar like project S8.

Based on the experimental result it is found that the local approach of hierarchical clustering algorithm takes less execution time space than global approach of hierarchical clustering algorithm. The number of iteration taken by hierarchical clustering local approach is also less than global approach of hierarchical clustering because links, goodness measure are also calculate to find the similarity among project. Hence the local approach performs better in terms of execution time and the number of iterations for small datasets.

REFERENCES

[1] Jiawei Han, Michelinekamber, Jian Pei, "Data Mining- Concepts and Techniques", Morgan kaufman Publishers, 3rd edition,2013.
[2] Arun K. Pujari, " Data Mining Techniques", University press , 3rd edition, 2013
[3] Parul Agarwal,M Afshar Alam, Ranjit Biswas ,Analysis the agglomerative hierarchical Clustering Algorithm for categorical Attributes- International journal of innovation, Management and Technology,Vol 1, No 2 june 2010 , ISSN:2010-0248 ,
[4] Sudipto Guha, Rajeev Rastogi and Kyuseok shim, " ROCK: A Robust Clustering Algorithm for Categorical Attributes", Proceedings of IEEE International Conference n Data Engineering, Sydney, March 1999.
[5] M Dutta, A. kakoti mahanta, Arun K. pujari, " QROCK: A Quick version of the ROCK Algorithm for Clustering of Categorical Data".
[6] IGNOU, "Project Guideline- Master of Computer Application", MCSP-060.
[7] Ashwina Tyagi, Sheetal Sharma, "Implementation of ROCK Clustering Algorithm For the optimization of Query searching time", International Journal on Computer Science and Engineering (IJCSE)., Vol 4, 2012.
[8] Guojun Gan, chaoqun ma, Jianhong Wu, "Data Clustering- Theory, algorithm and applications, ASA-SIAM, 2007.
[9] Poonam M. Bhagat, Prasad S. Halgaonkar,Vijay M.Wadhai, " Review of Clustering Algorithm for Categorical Data, International journal of Engineering and Advanced Technology(IJEAT), ISSN:2249-8958, Vol-3, Dec 2013.

AUTHOR PROFILE

Purnima Kumari Srivastava is working as a lecturer, Department of Information Technolohy, Ranchi Women's College, Ranchi, She completed her M.Tech in Computer Science from Birla Institute of technology Mesra Ranchi. Her two national papers has published..

Mamata Pandey is working as a lecturer, Department of Information Technolohy, Ranchi women's college, Ranchi, , She completed her M.Tech in Computer Science from Birla Institute of technology Mesra Ranchi. Her two national papers has published..

Vandana Bhattacherjee is working as a Professor, Department of Computer Science and Engineering, Birla Institute of Technology, Ranchi. She completed her B. E. (CSE) in 1989 and her M. Tech and Ph. D in Computer Science from JNU New Delhi in 1991 and 1995 respectively. She has over 80 National and International publications in Journal and Conference Proceedings. Her research areas include Software Process Models, Software Cost Estimation, Data Mining and Software Metrics.