Hierarchical Cluster Generation for Software Quality: A Comparative Approach

Jaya Pal[#1], Vandana Bhattacherjee[*2]

[#]Department of CS & E, Birla Institute of Technology
Ranchi, India
[1]jayapal@bitmesra.ac.in
[*]Department of CS & E, Birla Institute of Technology
Ranchi, India
[2]vbhattacharya@bitmesra.ac.in

Abstract— Clustering is a powerful technique of data mining for extracting useful information from a set of data and classifies the data into several clusters based on similarity of the pattern. This paper presents the quality estimation for students' projects data based on hierarchical clustering and fuzzy clustering using Min-Max method. From the experimental results it is seen the fuzzy clustering and hierarchical clustering technique prove to be useful tools in obtaining clusters which can be meaning fully interpreted.

Keyword- Hierarchical cluster, cluster generation, fuzzy clustering, comparative approach, software quality

I. Introduction

Clustering is a powerful data mining technique which groups the data into several clusters based on similarity of the pattern. The nature of the clusters may be either crisp or fuzzy. Fuzzy logic integrated with data mining techniques becomes one of the key constituents of soft computing in handling the challenges posed by massive collections of natural data [1]. The boundaries of the crisp clusters are well defined and fixed among themselves whereas fuzzy clusters have vague boundaries.

K-means is a popular clustering technique and its variations have proposed to overcome its inherent limitations [9] [10]. The clusters formed by K-means technique are crisp clusters. This technique has been used for software fault prediction [11]. Several methods of fuzzy clustering, such as Fuzzy C-Means [27], Fuzzy K-nearest neighborhood Algorithm [2], potential based clustering [3], Fuzzy clustering using max-min method [12] and others, have been proposed by various researchers.

The non-unique partitioning of the data in collection of clusters is the central idea in fuzzy clustering. The membership values of data points are assigned for each of the clusters. The membership value of zero indicates that the data point is not a member of the cluster under consideration. Handling of extreme outliers in many crisp techniques are difficult but the tendency of fuzzy clustering algorithms is to give them very small membership value in surrounding clusters [26].

The membership values with a maximum of one show the degree to which the data point represents a cluster. At the centre of the cluster, data points have maximum membership values and the membership value continuously decreases when we move away from the cluster's centre. Thus fuzzy clustering provides a flexible and robust method for handling natural data with vagueness and uncertainty [4]. In fuzzy clustering, for each cluster each data point will have an associated membership value. The membership value in the range [0,1] indicates the strength of association in that cluster. The compactness and distinctness of the clusters are decided based on the intra cluster and inter cluster distances of elements respectively.

Software Quality Estimation has been identified as one of the major challenges for computer science [5]. No method or model of estimation should be preferred over all others. Fuzzy logic may be used as a convenient tool for software development quality estimation[13][6].Soft computing technique likes fuzzy logic, case based reasoning have been used by several researchers for estimation of development cost and time in Software Engineering[7][14-18][19-23]. This research paper compares estimations obtained with simple cluster analysis method and fuzzy cluster analysis method. For this, three quality metrics have been gathered for 10 projects from 50 students working in 10 groups. These metrics are Graphics User Interface (GUI), Meaningful Error Message (MEM) and User Manual (UM). The rest of the paper is organized as follows: Section II presents the Cluster Analysis Method, Section III presents the Fuzzy Clustering Analysis, Section IV presents Illustration and Analysis, Section V presents Experimental Results and Section VI presents Conclusion.

II. CLUSTER ANALYSIS METHOD

Cluster analysis is a multivariate analysis technique where individual data points with similar characteristics are determined and grouped and dissimilar data points fall in different groups. The input of the cluster analytical system is a set of input data points and a standard of measuring the similarity between two data points. The output is a data set of several groups and these groups constitute a partition. An additional result of cluster analysis is the comprehensive description of each cluster, this result is particularly important for analysing the characteristics of collected data. There are various methods in cluster analysis such as systematic clustering (hierarchical clustering), fuzzy cluster, fuzzy cluster dynamic cluster etc.

In this paper, we have focused on hierarchical clustering which is essentially to combine the data points into clusters one by one, that is, N data points are considered as N clusters at first. For example, each data point becomes one cluster, then by calculating the distance between these clusters, and merging the two clusters with minimum distance into one cluster, we can get N – 1 clusters. By repeating the steps described above, at least one cluster must be merged each time until all the data points are incorporated in to one cluster [8]. For clustering n data points where each data point has m characteristics, using hierarchical clustering, each data point can be treated as a point in m dimensional Euclidian space. N data points are taken as n points of the m dimensional Euclidian space, these points form n X m matrix as

$$\begin{pmatrix} x_{11} & x_{12} & \dots\dots & x_{1m} \\ x_{21} & x_{22} & \dots\dots & x_{2m} \\ \dots & \dots & \dots\dots & \dots\dots \\ \dots & \dots & \dots\dots & \dots\dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix}$$

where x_{ij} represents the value of the j^{th} index of the i^{th} data point , j = 1,2, ….,m; i = 1,2,….., n. For clustering n data points the similarity of them is measured by calculating their distance. In clustering analysis Chebyshev distance, absolute value distance and Euclidian distance are used. Before using these distances the standardized transformation of original data should be performed and then distance could be calculated according to the changed data. There are various ways of standardized transformation such as mean-value regularization, extremely divergence regularization etc. In this paper, we take absolute value distance as the measure of distance to describe similarity among the data points and use mean-value regularization transformation for standardizing the original data.

The procedure of systematic (hierarchical) cluster analysis is as follows:

Step 1: *Perform standardize transformation of original data.*

Suppose there are n data points. Each of them is considered as one cluster, that is, there are n clusters. This paper uses mean- value regularization transformation for standardized transform.

For each j = 1,2,……,m,

$$\overline{x_j} = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

making the standardized transformation to the original data, let $x'_{ij} = \frac{x_{ij}}{\overline{x_j}}$, then obtain a new matrix as

$X' = \left(x'_{ij} \right)$,where i = 1,………,n; j = 1,……,m.

Step 2: *Calculate statistic of clusters i.e. shortest distance matrix.*

Shortest distance criteria in the hierarchical cluster analysis to measure the similarity of the two sub-clusters as

$$d_{ij} = \sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right|$$

Using this shortest distance construct the matrix of shortest distance as

$$\begin{bmatrix} 0 & & & & \\ d_{21} & 0 & & & \\ d_{31} & d_{32} & 0 & & \\ \ldots & \ldots & \ldots & \ldots & \\ d_{n1} & d_{n2} & \ldots & \ldots & 0 \end{bmatrix}_{nxn}$$

i.e $A^{(1)} = (d_{ij})_{nxn}$

where the elements in the matrix are the distance between two data points of all n data points. d_{ij} indicates the distance between C_i cluster and C_j cluster, $d_{ii} = 0$, since $d_{ij} = d_{ji}$, the matrix of distance is called symmetry matrix.

Step 3: *Clustering.*

Search for two clusters with minimum value in the matrix $A^{(1)}$ not including 0, merge the two clusters , then it reduces n-1 clusters. Again from the shortest distance matrix of n-1 clusters by calculating the distance among the new cluster and other n-2 clusters, and maintaining the distances of the remaining n-2 clusters. Then the matrix of distance is

$$A^{(2)} = \begin{bmatrix} 0 & & & & \\ d_{21} & 0 & & & \\ d_{31} & d_{32} & 0 & & \\ \ldots & \ldots & \ldots & & \\ d_{(n-1)1} & d_{(n-1)2} & \ldots & \ldots & 0 \end{bmatrix}_{(n-1) \times (n-1)}$$

Step 4: *Construct cluster analysis chart.*

There are several ways to represent the results of cluster analysis such as: tree graph, matrix table and pedigree diagram etc. This paper chooses matrix table and pedigree diagram to illustrate the results of cluster analysis.

III. FUZZY CLUSTERING ANALYSIS

The fuzzy cluster analysis approach makes use of fuzzy equivalent relation to classify the objects into different criterion [24]. Before introducing fuzzy cluster analysis, we need to know two relations, fuzzy similarity relation and fuzzy equivalent relation.

A. Fuzzy similar relation and fuzzy equivalent relation

1) Fuzzy similar relation: Let $R = (r_{ij})_{nxn}$ is a fuzzy relation on U, which satisfies the following conditions: (a) Reflexivity: $r_{ii} = 1$; (b) Symmetry: $r_{ij} = r_{ji}$, then $R = (r_{ij})_{nxn}$ is a fuzzy relation.

2) Fuzzy equivalent relation: Let $R = (r_{ij})_{nxn}$ is a fuzzy relation on U, which satisfies the following conditions: (1) Reflexivity: $r_{ii} = 1$; (2) Symmetry: $r_{ij} = r_{ji}$; (3) Transitivity $R \circ R \subseteq R$, then $R = (r_{ij})_{nxn}$ is a equivalent relation.

Step1: *Selecting attributes of fuzzy cluster analysis.*

Different attributes should be selected for analysis, for example, in this paper when analyzing software usability, we select the user manual (UM), graphical user interface(GUI) and meaningful error messages(MEM) attributes.

Step 2: *Standardizing the data.*

Let P1, P2,……Pn indicate data points having m number of attributes then we use formula for standardization as

$$P_{ij} = \frac{(P_{ij} - \min(P_j))}{(\max(P_j) - \min(P_j))} \qquad \ldots\ldots (1)$$

where $1 \leq i \leq n$ and $1 \leq j \leq m$

Step3: *Calculating fuzzy similarity matrix.*

There are various methods to calculate fuzzy similarity matrix. Some of them are as follows:

(i)Euclidean distance method

$$r_{ij} = \sqrt{\frac{1}{n}\sum_{k=1}^{m}(x_{ik} - x_{jk})^2} \qquad \text{...... (2)}$$

where x_{ik} is the value of point i and number k and x_{jk} is the value of point j and factor number k.

(ii) Correlation coefficient method

$$r_{ij} = \frac{\sum_{k=1}^{m}\left(x_{ik} - \overline{x_i}\right)\left(x_{jk} - \overline{x_j}\right)}{\sqrt{\sum_{k=1}^{m}\left(x_{ik} - \overline{x_i}\right)^2} \times \sqrt{\sum_{k=1}^{m}\left(x_{jk} - \overline{x_j}\right)}^2} \qquad \text{...... (3)}$$

where $\overline{x_i} = \frac{1}{n}\sum_{k=1}^{m} x_{ik}$, $\overline{x_j} = \frac{1}{n}\sum_{k=1}^{m} x_{jk}$

(iii) Minimum-Maximum method

$$r_{ij} = \frac{\sum_{k=1}^{m}\min(x_{ik,}\ x_{jk})}{\sum_{k=1}^{m}\max(x_{ik}\ , x_{jk})} \qquad \text{...... (4)}$$

where x_{ik} is the value of point i and factor number k, and x_{jk} is the value of point j and factor number k.

Step4: *Clustering based on the fuzzy similarity matrix.*

We use the method of transitive closure to obtain fuzzy equivalent matrix [10] as follows:

A fuzzy relation which has symmetry, reflexivity and transitivity is called a matrix of equivalence relation. For clustering, fuzzy equivalence matrix can be obtained by several composition computations. The procedure is: determine $R^2 = R \circ R$, determine $R^4 = R^2 \circ R^2$,………., determine $R^{2k} = R^k$ and stop, R^k is just fuzzy equivalence relation. i.e. the transitive closure t(R) of R equals to R^k.But it is inconvenient to calculate the fuzzy equivalent matrix when order of matrix is high. So we use the method of direct clustering based on the similar fuzzy relation to simplify the calculation. The method of direct cluster uses fuzzy similar matrix to calculate the result. Its principle of clustering is: x_i and x_j are the same on level λ if and only if the fuzzy similarity matrix has a route connecting x_i and x_j whose weight is not smaller than λ.

Step5: *Constructing Dendrogram cluster graph*.

IV. ILLUSTRATION AND ANALYSIS

In this paper we have used Systematic cluster analysis and Fuzzy cluster analysis to classify the software projects on the basis of software quality and compare both the methods. Objects collection consists of 10 projects: U= {x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}}as shown in Table 1.

TABLE 1
Projects and metrics: Graphical user interface (GUI), Meaningful Error Message (MEM), User Manual (UM), Software Quality (SQ) (ranks)

Project	GUI	MEM	UM	SQ
P1	0	0.5	9	75
P2	5	0.5	14	80
P3	1	0.4	8	72
P4	7	0.7	12	82
P5	7	0.7	16	82
P6	6	0.6	14	83
P7	7	0.8	18	91
P8	1	0.2	9	62
P9	7	0.5	14	82
P10	8	0.8	17	92

Statistical analysis of above data

	GUI	**MEM**	**UM**	**SQ**
Min	0	0.2	8	62
Max	8	0.8	18	92
Max-Min	8	0.6	10	30

The extracted data above is utilized to realize cluster analysis using systematic cluster analysis and fuzzy cluster analysis. Each project is considered as one cluster.

A. Metrics Used

This paper focuses on quality of software using clustering and metrics were designed and / or adapted from Pal and Bhattacherjee [25] where the authors have developed a Fuzzy Logic System for prediction of software quality.

Description of metrics:

1) *GUI (Graphical User Interface):* GUI was measured as the relative number of forms which were clearly displayed, on a scale of 0-10.
2) *MEM (Meaningful Error Message):* MEM was measured as the relative number of meaningful error messages displayed by the software, on a scale of 0-1.
3) *UM (User Manual):* UM was measured as the completeness of the user manual or help file, on a scale of 1-20.

The usability of the ultimate product (program) has been judged by team of three experts who ranked the various projects on a scale of 50-100 for usability and this served as the predicted output

B. Application of Systematic Cluster Analysis to Software Quality

In which each project is regarded as one cluster first, then merging most similar clusters into a new sub-cluster, and combining the new sub-cluster with other clusters further according to similarity of them. The step would continue until all sub- clusters merge into one cluster.

The procedure is as follows**:**

Step 1: *Standardize transformation of original data.*

Each project data point as shown in Table 1 is treated as one point of 3-Dimensional Euclidean space, and 10 projects are viewed as 10 points of 3- Dimensional Euclidean space, which forms 10 * 3 matrix.

$$A = \begin{pmatrix} 0 & 5 & 1 & 7 & 7 & 6 & 7 & 1 & 7 & 8 \\ 0.5 & 0.5 & 0.4 & 0.7 & 0.7 & 0.6 & 0.8 & 0.2 & 0.5 & 0.8 \\ 9 & 14 & 8 & 12 & 16 & 14 & 18 & 9 & 14 & 17 \end{pmatrix}^T$$

Ten projects could be represented as 10 clusters.

that is C1= (0, 0.5, 9), C2 = (5, 0.5, 14), C3 = (1, 0.4, 8), C4 = (7, 0.7, 12), C5 = (7, 0.7, 16), C6 = (6, 0.6, 14), C7 = (7, 0.8, 18), C8 = (1, 0.2, 9), C9 = (7, 0.5, 14), C10 = (8, 0.8,17)

Standardize the matrix and calculate

$$\overline{x_j} = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

That is $\overline{X_1} = 4.9$, $\overline{X_2} = 0.57$, $\overline{X}_3 = 13.1$

From $x_{ij}^{'} = \frac{x_{ij}}{\overline{x_j}}$, we get standardized matrix as

$$\begin{pmatrix} 0.0 & 0.87 & 0.68 \\ 1..02 & 0.87 & 1.07 \\ 0.20 & 0.70 & 0.61 \\ 1.42 & 1.23 & 0.92 \\ 1.42 & 1.23 & 1.22 \\ 1.22 & 1.05 & 1.06 \\ 1.42 & 1.40 & 1.37 \\ 0.20 & 0.35 & 0.69 \\ 1.42 & 0.85 & 1.07 \\ 1.63 & 1.40 & 1.29 \end{pmatrix}$$

Step 2 and Step 3: *Construction of shortest distance matrix and clustering.*

Using the shortest distance, $d_{ij} = \sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right|$ construct the shortest distance matrix as follows:

$$A^{(1)} = \begin{pmatrix} 0 & 1.41 & 0.44 & 2.02 & 2.32 & 1.78 & 2.64 & 0.73 & 1.83 & 2.77 \\ 1.41 & 0 & 1.45 & 0.91 & 0.91 & 0.39 & 1.23 & 1.72 & 0.42 & 1.36 \\ 0.44 & 1.45 & 0 & 2.06 & 2.36 & 1.82 & 2.68 & 0.43 & 1.83 & 2.81 \\ 2.02 & 0.91 & 2.06 & 0 & 0.30 & 0.52 & 0.62 & 2.33 & 0.53 & 0.75 \\ 2.32 & 0.91 & 2.36 & 0.30 & 0 & 0.54 & 0.32 & 2.63 & 0.53 & 0.45 \\ 1.78 & 0.39 & 1.82 & 0.52 & 0.54 & 0 & 0.86 & 2.09 & 0.41 & 0.99 \\ 2.64 & 1.23 & 2.68 & 0.62 & 0.32 & 0.86 & 0 & 2.95 & 0.75 & 0.29 \\ 0.73 & 1.72 & 0.43 & 2.33 & 2.63 & 2.09 & 2.95 & 0 & 2.10 & 3.08 \\ 1.83 & 0.42 & 1.83 & 0.53 & 0.53 & 0.41 & 0.75 & 2.10 & 0 & 0.98 \\ 2.77 & 1.36 & 2.81 & 0.75 & 0.45 & 0.99 & 0.29 & 3.08 & 0.98 & 0 \end{pmatrix}$$

In the matrix A $^{(1)}$, the shortest distance is $d^{7,10} = 0.29$ and the level of aggregation is 0.29. By merging clusters C7 and C10 into new cluster C11, we get nine sub-clusters {C11, C1, C2, C3, C4, C5, C6, C8, C9}.Standardized Mean of C11 calculated from standardized matrix = (1.52, 1.40, 1.33).Therefore standardized transformation of remaining (excluding C7 and C10) original data including C11 is as follows:

				Order of clusters
C11	1.52	1.40	1.33	1
C1	0	0.87	0.68	2
C2	1.02	0.87	1.07	3
C3	0.20	0.70	0.61	4
C4	1.42	1.23	0.92	5
C5	1.42	1.23	1.22	6
C6	1.22	1.05	1.06	7
C8	0.20	0.35	0.69	8
C9	1.42	0.85	1.07	9

Now the new shortest distance matrix in terms of nine sub-clusters as shown above is as follows:

$$A^{(2)} = \begin{pmatrix} 0 & 2.70 & 1.29 & 2.74 & 0.68 & 0.38 & 0.92 & 3.01 & 0.91 \\ 2.70 & 0 & 1.41 & 0.44 & 2.02 & 2.32 & 1.78 & 0.73 & 1.83 \\ 1.29 & 1.41 & 0 & 1.45 & 0.91 & 0.91 & 0.39 & 1.72 & 0.42 \\ 2.74 & 0.44 & 1.45 & 0 & 2.06 & 2.36 & 1.82 & 0.43 & 1.83 \\ 0.68 & 2.02 & 0.91 & 2.06 & 0 & 0.30 & 0.52 & 2.33 & 0.53 \\ 0.38 & 2.32 & 0.91 & 2.36 & 0.30 & 0 & 0.54 & 2.63 & 0.53 \\ 0.92 & 1.78 & 0.39 & 1.82 & 0.52 & 0.54 & 0 & 2.09 & 0.41 \\ 3.01 & 0.73 & 1.72 & 1.43 & 2.33 & 2.63 & 2.09 & 0 & 2.10 \\ 0.91 & 1.83 & 0.42 & 1.83 & 0.53 & 0.53 & 0.41 & 2.10 & 0 \end{pmatrix}$$

According to $A^{(2)}$, the shortest distance is $d^{5,6} = 0.30$ .By merging clusters C4 and C5 into new cluster C12, we get new shortest distance matrix $A^{(3)}$ in terms of eight sub-clusters { C12,C11,C1,C2,C3,C6,C8,C9 },and so on. The clustering process will continue until

$$A^{(9)} = \begin{pmatrix} 0 & 2.21 \\ 2.21 & 0 \end{pmatrix}$$

According to $A^{(9)}$, the shortest distance is $d^{1,2} = 2.21$.By merging clusters C18 and C16 we get new clusters C19 contains all data points. The order of clustering is shown in Table 2.

TABLE 2
Clustering order

Combined order	**Combined clusters**	**Level of aggregation**
1	C11={C7,C10}	0.29
2	C12 = {C4,C5}	0.30
3	C13 = {C12,C9}	0.38
4	C14 = {C13,C6}	0.40
5	C15 = {C3,C8}	0.43
6	C16 = {C15,C1}	0.57
7	C17 = {C14,C2}	0.58
8	C18 = {C11,C17}	0.83
9	C19 = {C16,C18}	2.21

Step 4: *Obtain Cluster analysis chart.*
Dendrogram or hierarchical cluster diagram is shown in Fig.1.

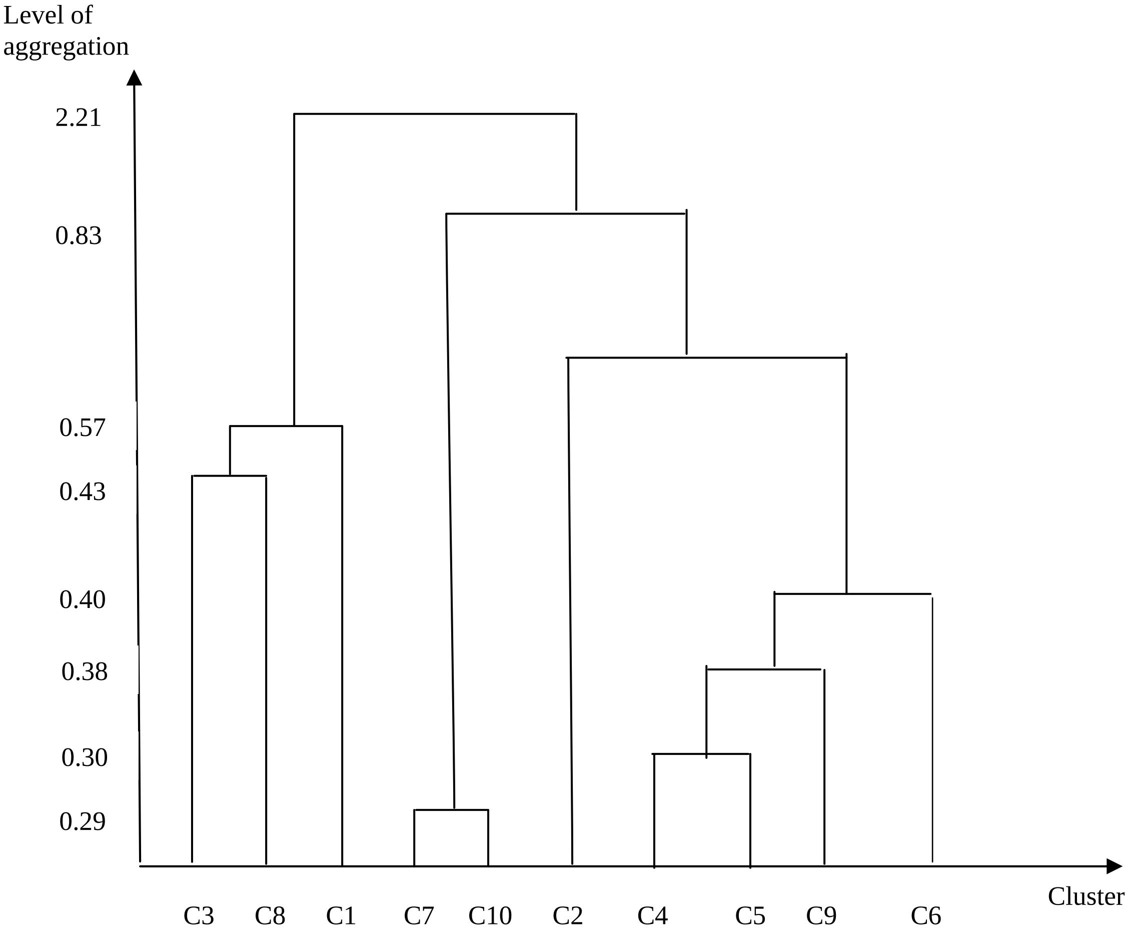

Fig 1: Dendrogram for hierarchical clustering

1) Analysis of Systematic (Simple) Clustering: For the hierarchical clustering analysis, it is noted from matrix $A^{(1)}$ that the shortest distance is 0.29 hence data points 7 and 10 are merged into cluster C11. The next shortest distance obtained is 0.30 and data points 4 and 5 get merged to get C12. Proceeding in this manner, data point 9 gets merged with cluster C12 (4 and 5) at 0.38 to get C13, data point 6 gets merged with C13 (4, 5 and 9) at 0.40 to get C14, data points 3 and 8 merged together at 0.43 to obtain C15, data point 1 merges with C15 (3 and 8) at 0.57 to get C16 and data point2 merges with C14 (4, 5, 6 and 9)at 0.58 to get C17. At 0.83, C11 and C17 merge together to get C18 (4, 5, 6, 9, 2, 7 and 10) and at 2.21 C18 and C16 merge to obtain one final cluster C19. The dendrogram for hierarchical clustering analysis is shown in Fig 1.

C. Application of Fuzzy Cluster Analysis to Software Usability

The steps of fuzzy clustering are as follows:

Step1: *Selecting attributes of fuzzy cluster analysis.*

Different attributes selected for analysis are user graphical user interface (GUI, user manual (UM) and meaningful error messages (MEM) as shown in Table 1.

Step2: *Standardizing the data in Table1.*

On using equation (1) for standardization, we get the standardized matrix as

$$\begin{pmatrix} 0.0 & 0.50 & 0.10 \\ 0.62 & 0.50 & 0.60 \\ 0.12 & 0.33 & 0.0 \\ 0.87 & 0.83 & 0.40 \\ 0.87 & 0.83 & 0.80 \\ 0.75 & 0.66 & 0.60 \\ 0.87 & 1.0 & 1.0 \\ 0.12 & 0.0 & 0.10 \\ 0.87 & 0.50 & 0.60 \\ 1.0 & 1.0 & 0.90 \end{pmatrix}$$

Step 3: *Construction of fuzzy similarity matrix.*

We have used Minimum- Maximum method to calculate fuzzy similarity matrix $R = (r_{ij})_{10x10}$ on using formula (4) and data in Table 1.

$$r_{ij} = \frac{\sum_{k=1}^{m} \min(x_{ik}, x_{jk})}{\sum_{k=1}^{m} \max(x_{ik}, x_{jk})}$$

where $r_{ij} \in [0,1]$

We get fuzzy similarity matrix as follows:

$$A_{\min-\max} = \begin{pmatrix} 1.0 & 0.35 & 0.46 & 0.28 & 0.24 & 0.29 & 0.21 & 0.14 & 0.30 & 0.21 \\ 0.35 & 1.0 & 0.26 & 0.66 & 0.68 & 0.85 & 0.60 & 0.13 & 0.87 & 0.59 \\ 0.46 & 0.26 & 1.0 & 0.21 & 0.18 & 0.22 & 0.16 & 0.22 & 0.22 & 0.15 \\ 0.28 & 0.66 & 0.21 & 1.0 & 0.84 & 0.78 & 0.73 & 0.10 & 0.76 & 0.72 \\ 0.24 & 0.68 & 0.18 & 0.84 & 1.0 & 0.80 & 0.87 & 0.08 & 0.78 & 0.86 \\ 0.29 & 0.85 & 0.22 & 0.78 & 0.80 & 1.0 & 0.70 & 0.11 & 0.86 & 0.69 \\ 0.21 & 0.60 & 0.16 & 0.73 & 0.87 & 0.70 & 1.0 & 0.07 & 0.68 & 0.92 \\ 0.14 & 0.13 & 0.22. & 0.10 & 0.08 & 011 & 0.07 & 1.0 & 0.11 & 0.07 \\ 0.30 & 0.87 & 0.22 & 0.76 & 0.78 & 0.86 & 0.68 & 0.11 & 1.0 & 0.67 \\ 0.21 & 0.59 & 0.15 & 0.72 & 0.86 & 0.69 & 0.92 & 0.07 & 0.67 & 1.0 \end{pmatrix}$$

Step 4: *Clustering based on the fuzzy similarity matrix.*

From similarity matrix A $_{\text{min-max}}$,

When λ = 1.0, 10 sub clusters are divided.

{C1}, {C2},{C3}, {C4},{C5,}{C6},{C7},{C8},{C9},{C10}

When λ = 0.92, 9 sub clusters are divided.

{C7, C10}, {C1}, {C2}, {C3}, {C4}, {C5}, {C6}, {C8}, {C9}

When λ = 0.87, 8 sub clusters are divided.

{C7, C10}, {C2, C9}, {C1}, {C2}, {C3}, {C4}, {C5}, {C6}, {C8}

When λ = 0.86, 7 sub clusters are divided.

{C2, C9, C6}, {C7, C10}, {C1}, {C3}, {C4}, {C8}, {C5}

When λ = 0.84, 6 sub clusters are divided.

{C2, C9, C6}, {C7, C10}, {C4,C5}, {C1},{C3}, {C8}

When λ = 0.80, 5 sub clusters are divided.

{C2, C9, C6, C4, C5}, {C7, C10}, {C1}, {C3}, {C8}

When λ = 0.73, 4 sub clusters are divided.
{C2, C4, C5, C6, C7, C9, C10}, {C1}, {C3}, {C8}
When λ = 0.46, 3 sub clusters are divided.
{C2, C4, C5, C6, C7, C9, C10}, {C1, C8}, {C3}
When λ = 0.22, 2 sub clusters are divided.
{C2, C4, C5, C6, C7, C9, C10}, {C1, C3, C8}
When λ = 0.21, 1 cluster is obtained.
{C1, C2, C3, C4, C5, C6, C7, C8, C9, C10}
The order of clustering is shown in Table 3.

TABLE 3
Clustering order

Combined order	**Combined clusters**	**Level of similarity measures**
1	C11={C7,C10}	0.92
2	C12 = {C2,C9}	0.87
3	C13 = {C12,C6}	0.86
4	C14 = {C4,C5}	0.84
5	C15 = {C13,C14}	0.80
6	C16 = {C11,C15}	0.73
7	C17 = {C1,C3}	0.46
8	C18 = {C17,C18}	0.22
9	C19 = {C16,C18}	0.21

Step 5: Obtain Dendrogram cluster graph.
Dendrogram or hierarchical cluster diagram is shown in Fig. 2.

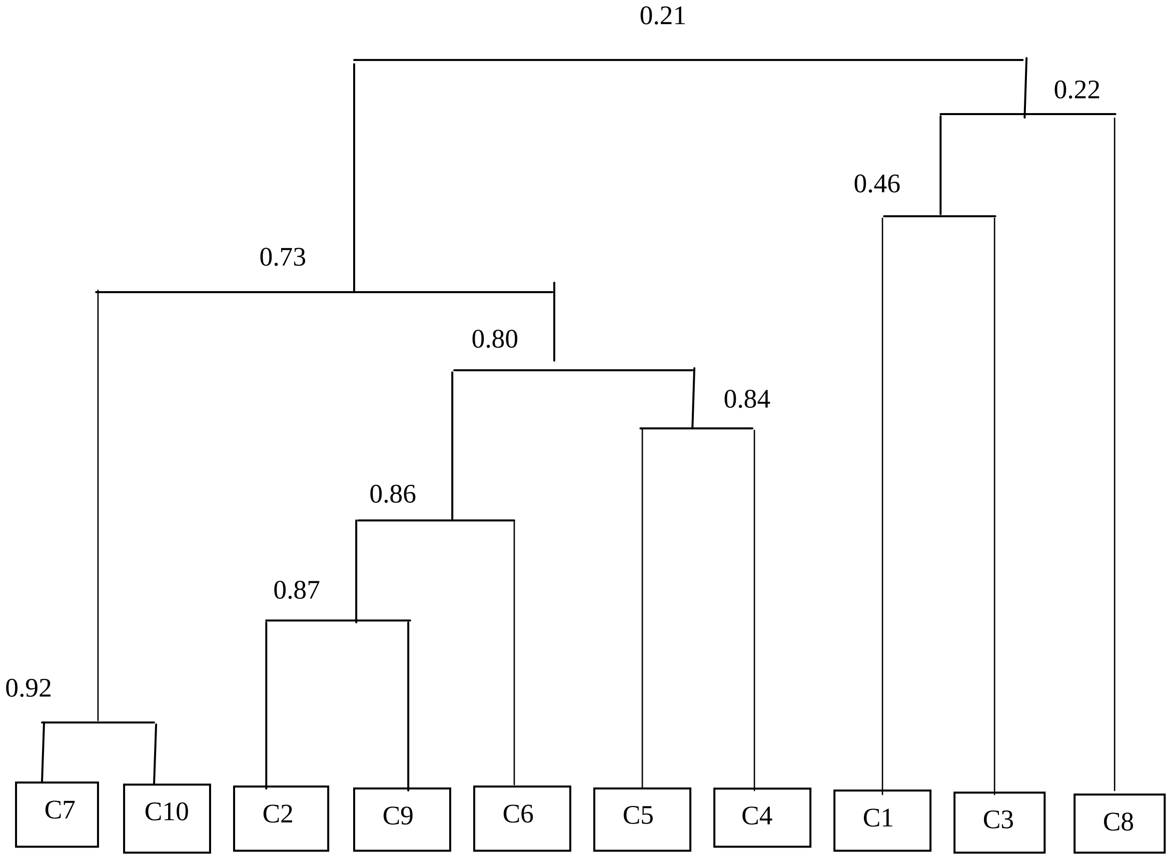

Fig 2.Dendrogram for fuzzy hierarchical clustering

1) *Analysis of fuzzy clustering* : After standardization of data, the similarity matrix A $_{\text{min-max}}$ is obtained by Min-Max method. For the hierarchical clustering analysis, it is noted from matrix A $_{\text{min-max}}$ that at λ = 1.0, 10 sub clusters are divided ie each data point is in its own cluster. At λ = 0.92, data points 7 and 10 combine to get cluster C11. Proceeding in this manner, data points 2 and 9 get merged together at λ = 0.87 to obtain C12, data point 6 gets merged with C12 (C2 and C9)at λ = 0.86 to obtain C13,data points 4 and 5 merge to get C14 at λ = 0.84. At λ = 0.80, clusters C13 and C14 are merged to obtain C15 (C2, C9, C6, C4 and C5), at λ = 0.73, clusters C15 and C11 are merged to obtain C16 (C2, C9, C6, C4, C5, C7 and C10) at λ = 0.46, data points 1 and 3 merge to get C17, at λ = 0.22, data point 8 is merged to get C18 (C1,C3 and C8) and at λ = 0.21, C16 and C18 merge to obtain one final clusterC19(C1, C2, C3, C4, C5, C6, C7, C8, C9, and C10). The dendrogram for hierarchical clustering analysis is shown in Fig 2.

V. EXPERIMENTAL RESULTS

The sum squared error (SSE) using Euclidian distance for combined clusters of simple cluster analysis method and fuzzy cluster analysis method is calculated using the following formula:

$$\text{SSE} = \sum_{i=1}^{k} \sum_{j=1}^{j=|c_i|} \left\| x_{j \in c_i} - \overline{x_i} \right\|$$

Where, $|c_i|$ is the cardinality of i th cluster.

k is the number of combined clusters of each method.

$\overline{x_i}$ is mean of i th cluster.

The sum squared error(SSE) of combined clusters (Simple cluster analysis method) as shown in Table 2 is depicted in the following Table 4.

TABLE 4
Sum squared error of clusters in Simple cluster analysis method

Level of aggregation	Combined Clusters	Number of clusters	Sum Squared error
0.29	{C7,C10}	9	0.224
0.30	{C7,C10} {C4,C5}	8	0.524
0.38	{C7,C10}{C4,C5,C9}	7	0.870
0.40	{C7,C10} (C4,C5,C9,C6}	6	1.046
0.43	{C7,C10}{C4,C5,C9,C6}{C3,C8}	5	1.404
0.57	{C7,C10}{C4,C5,C9,C6}{C3,C8,C1}	4	1.713
0.58	{C7,C10}{C2,C4,C5,C9,C6}{C3,C8,C1}	3	2.062
0.83	{C7,C10,C2,C4,C5,C9,C6}{C3,C8,C1}	2	2.728
2.21	{C7,C10,C2,C4,C5,C9,C6,C3,C8,C1}	1	6.438

The Sum squared error of combined clusters (Fuzzy cluster analysis method) as shown in Table 3 is depicted in the following Table 5.

Table 5

Sum squared error of clusters in Fuzzy cluster analysis method

Level of similarity measure	Combined Clusters	Number of clusters	Sum Squared error
0.92	{C7,C10}	9	0.164
0.87	{C7,C10} {C2,C9}	8	0.414
0.86	{C7,C10}{C2,C9,C6}	7	0.534
0.84	{C7,C10} (C2,C9,C6}{C4,C5}	6	0.934
0.80	{C7,C10}{C2,C9,C6,C5,C4}	5	1.160
0.73	{C7,C10,C2,C9,C6,C5,C4}	4	2.107
0.46	{C7,C10,C2,C9,C6,C5,C4}{C1,C3}	3	2.337
0.22	{C7,C10,C2,C9,C6,C5,C4}{C1,C3,C8}	2	2.721
0.21	{C7,C10,C2,C4,C5,C9,C6,C3,C8,C1}	1	4.909

The comparison results of both the methods are depicted in Table 6.

TABLE 6
Comparison results

Methods	MSSE
Simple clustering method	1.701
Fuzzy clustering method	1.528

From this result, it is observed that fuzzy clustering method yields best clusters.

VI. CONCLUSION

This paper presents the quality estimation for students' projects data based on hierarchical clustering and fuzzy clustering using Min-Max method. In the hierarchical clustering approach the first step is to perform standardized transformation of original data. We then compute shortest distance matrix, using the minimum value in this matrix we merge two clusters at a time iteratively until one cluster for the entire data remains. Results of this cluster analysis are presented in Fig.1. Fuzzy clustering analysis makes use of fuzzy similarity matrix to classify the objects into different criterion. First attributes for fuzzy cluster analysis are selected and their values are standardized. Using these, fuzzy similarity matrix is calculated. We use the Min-Max method for this purpose (equation 4) then the fuzzy similarity matrix is computed. From this matrix, based on different threshold values the various clusters are obtained. Finally the cluster analysis chart is obtained and presented in Fig.2.

From the experimental results presented in Tables 4 and 5 of Section V, it is seen that for obtaining three clusters, the SSE of hierarchical clustering approach is smaller and gives the clusters as{C7, C10}{C2, C4,C5, C9, C6}{C3, C8,C1} whereas for fuzzy clustering the clusters obtained are{C7,C10,C2,C9,C6,C5,C4}{C1,C3}.For obtaining two clusters the SSE of fuzzy clustering is smaller and clusters obtained are {C7,C10,C2,C9,C6,C5,C4}{C1,C3,C8}whereas for hierarchical clusters obtained are {C7,C10,C2,C4,C5,C9,C6}{C3,C8,C1}. If the students are to be grouped into two categories the fuzzy clustering gives better results whereas for three categories hierarchical clustering gives lesser SSE values. As part of our ongoing work, we are collecting exhaustive sets of data so as to develop a model which can be for generalized use. Future research involves in collecting more data to serve as the basis of a generalized fuzzy tool for quality prediction.

REFERENCES

[1] S. K.Pal, and P. Mitra, "Data Mining in Soft Computing Framework: A survey," IEEE transactions on neural networks, vol13, no1,January 2002.

[2] J. Keller, M. R. Gray, and Givens, J.A. "A Fuzzy K-nearest neighbor algorithm," IEEE Trans. on Systems, Man and Cybernetics, SMC-15, 4, PP. 580-585, 1985.

[3] S. L. Chiu, "Fuzzy model identification based on cluster estimation," Journal of Intelligent Fuzzy Systems, 2, pp.267-278, 1994.

[4] B. Thomas , G. Raju , and S. Wangmo, "A modified Fuzzy C-Means Algorithm for Natural Data Exploration," World Academy of Science, Engineering and Technology 49,2009.

[5] P. Brooks Fredrick, Jr., " Three Great Challenges for Half-Century-Old Computer Science," Journal of the ACM, Vol. 50, No. 1 pp. 25-26, January 2003.

[6] L.C.Briand, I. Wieczorek, "Software Resource Estimation Encyclopedia of Software Engineering," Volume 2, New York: John Wiley & Sons, pp. 1160-1196.

[7] P. Musflek , W. Pedrycz, G. Succi and M. Reformat , "Software Cost Estimation with Fuzzy Models," Applied Computing Review, Vol. 8, No. 2, 2000, pages 24-29.

[8] Li Dong, Xie Zong-bao, and Zheng Qiu-yan, Discussion on cluster analysis-based grouping method of collaborative learning. Software Guide, 2005, (6): 28-30.
[9] P.S. Bishnu, and V. Bhattacherjee, “A new Initialization Method for K-Means Algorithm using Quad Tree,” NCM2C, 2008, JNU, New Delhi.
[10] P.S. Bishnu, and V. Bhattacherjee, “Divide and conquer based clustering using K-Means algorithm,” Conference on Information Science, Technology and Management, 2009.
[11] P.S. Bishnu, and V. Bhattacherjee, “Unsupervised learning approach to fault prediction in Software Module,” in proceedings of National Conference on Computing and Systems 2010, Burdwan, India,January2010, pp 91-94, 2010.
[12] J. Pal, and V. Bhattacherjee, “Application of fuzzy clustering on software quality using max-min method ,” in proceedings of International Conference on Communication, Computing & Security ,NIT , Rourkela, October 2012.
[13] Witold Pedrycz, “Computational Intelligence as an Emerging paradigm of Software Engineering,” SEKE 2002, Ischia, Italy.
[14] X. Huang, J. Ren and L.F. Capretz, “A Neuro-Fuzzy Tool for Software Estimation ,” Proceedings of the 20th IEEE International Conference on Software Maintenance, pp. 520, 2004.
[15] Idri, A. Abran, and T. Khoshgoftaar, “Fuzzy Analogy: a New Approach for Software Cost Estimation,” International Workshop on Software Measurement (IWSM´01), Montréal, Québec, Canada, August 28-29, 2001.
[16] Idri, A. Abran, and T. Khoshgoftaar, “Estimating Software Project Effort by Analogy Based on linguistic Values,” Proceedings of the Eighth IEEE Symposium on Software Metrics (METRICS.02)2002.
[17] Gray and S. MacDonell, “Applications of Fuzzy logic to Software Metric Models for Development Effort Estimation,” 1997.
[18] M. Braz, and S. Vergilio, “Using Fuzzy Theory for Effort Estimation of Object-Oriented Software,” Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2004.
[19] V. Bhattacherjee and S. Kumar, “An Expert Case Based Framework for Software Cost Estimation,” in proceedings SCT-2006, pp 183-188.
[20] V. Bhattacherjee , “The Soft Computing Approach to Program Development Time,” in proceedings ICTT 2006, pp.291-292.
[21] N. Vira and V. Bhattacherjee, “Software Development Time Estimation Using Fuzzy Logic,” presented at TECHTATVA 2009,Manipal Institute of Technology, Karnataka, India, September 8-11, 2009.
[22] V. Bhattacherjee, S. Kumar and E. Rashid, “A Cases Study on Estimation of Software Development Effort,” in proceedings ICACT, 2008.
[23] V. Bhattacherjee, S. Kumar and E. Rashid, “Case Based Estimation Model Using Project Feature Weight,” in proceedings RAIT 2009, pp 142-246.
[24] Qi Yang, “Application of Fuzzy Cluster Analysis to Tax Planning for Location of Foreign Direct Investment,” IEEE 3rd International Conference on Information Management, Innovation Management and Industrial Engineering, PP. 553-555, 2010.
[25] J. Pal, and V. Bhattacherjee, “A Fuzzy Logic System for Software Quality Estimation,” in proceedings ICIT 2009, pp 183-187, 2009.
[26] G. Carl Looney,“A Fuzzy clustering and Fuzzy Merging Algorithm”, Available: http://citeseer.ist.psu.edu/399498.html
[27] J. C. Bezdek, “Fuzzy mathematics in pattern classification,” Ph.D.thesis, Applied Mathematics Centre, Cornell University, Ithaca, 1973.

AUTHOR PROFILE

Jaya Pal is working as Assistant Professor, Department of Computer Science & Engineering, Birla Institute of Technology, Ranchi, India. She received her Master in Computer Science Application from Birla Institute of Technology, Mesra, Ranchi, India in 2003. Her research area includes Data Mining and Soft Computing.

Vandana Bhattacherjee is working as Professor, Department of Computer Science & Engineering, Birla Institute of Technology, Ranchi, India. She completed her B.E. (CSE) in 1989 and her M. Tech and Ph.D. in Computer Science from JNU, New Delhi, India in 1991 and 1995 respectively. She has over 65 National and International publications in Journal and Conference Proceedings. She is a member of IEEE Computer Society of India. Her research area include Software Process Models, Software cost estimation, Data Mining and Software Metrices and Soft Computing.