ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 3, September 2012

Data Analysis of Bio-Medical Data Mining using Enhanced Hierarchical Agglomerative Clustering

V.V.Jaya Rama Krishnaiah, D.V.Chandra Sekar, Dr. K.Ramchand H Rao

Abstract: ***Data mining became increasingly important in bioinformatics and biomedical area during last decade. Various data mining methods, like clustering, have successfully revealed previous unknown knowledge in bioinformatics and biomedical area. Hierarchical Clustering is one of the wide areas of knowledge analysis, within the categorical data mining domain. In several contexts and domains, hierarchical agglomerative clustering (HAC) offers best-quality results, but at the price of a high complexity which reduces the size of datasets which can be handled. In some contexts, in particular, computing distances between objects is the most expensive task. In this paper we propose an approach called Enhanced Hierarchical Clustering Approach (EHAC), aimed at improving performance, reliability of data, which is well integrated in all the phases of the Entropy based Mean Clustering Approaches and can be applied to single-linkage HAC Process. After describing the method, we provide some theoretical evidence of its pruning power, followed by an empirical study of its effectiveness over different data domains, with a special focus on dimensionality issues.***

Keywords: **Data mining, Agglomerative Hierarchical Clustering (AHC), Entropy Based Mean (EBM) Clustering, Dendrogram, Euclidean Distances, Heart attack.**

I. INTRODUCTION

Cluster analysis is a tool of exploratory data analysis that tries to find the intrinsic structure of data by organizing patterns into groups of clusters. It is important for clustering how to define a cluster and how to evaluate clustering results. The clustering problems can be divided into attributive and non-attributive according to presence of absence of any properties of the objects considered. In an attributive case, selected properties of the objects are utilized in the clustering process. In non-attributive case, non properties are known except the distance between the objects. There are a number of different methods that can be used to carry out a cluster analysis; these methods can be classified as Hierarchical methods and Non-Hierarchical Methods In hierarchical clustering, objects are linked to each other in a hierarchical way. Hierarchical methods distinguished into Agglomerative and Divisive Methods. In Agglomerative methods, in which subjects start in their own separate cluster. The two 'closest' (most similar) clusters are then combined and this is done repeatedly until all subjects are in one cluster. At the end, the optimum number of clusters is then chosen out of all cluster solutions. In Divisive methods, in which all subjects start in the same cluster and the above strategy is applied in reverse until every subject is in a separate cluster. Agglomerative methods are used more often than divisive methods, so this handout will concentrate on the former rather than the latter. In Agglomerative Hierarchical Clustering, similarity or distance measure has to be defined for this technique. When two objects are linked to each other, the distances of this pair of objects to all other objects, not yet linked to this pair by previous links, are modified. Define the distance between two objects O_i and O_j as being d (O_i,O_j). When the objects O_i and O_j are linked (denoted by object O_{ij}), the distance of an object O_k to this pair O_{ij} is in general given by:

$$d(O_k,O_{ij}) = d(O_k,O_j)+d(O_k,O_j)+d(O_k,O_j)-d(O_k,O_j)$$

Final clusters are defined by placing a threshold on the values of d(O_i,O_j), at which (groups of) objects are linked, taking into account the structure of the dendrogram. A *Dendrogram is* a branching diagram that hierarchically nests objects into increasingly more inclusive groups, degree of similarity is depicted by length of branch, ordering axis is to prevent branches from crossing but is otherwise arbitrary. In Non-hierarchical methods, often known as k-means and Entropy based Mean Clustering methods, is one of the simplest unsupervised approach, defines the clusters with respect to predefined number of clusters. These clusters were defined through the defined object function, in many cases it is the squared error function, like

Object function $J = \sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^{(j)} - C_j\right\|^2$

Where $\left\|x_i^{(j)} - C_j\right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre C_j is an indicator of the distance of the n data points from their respective cluster centres.

II. PROPOSED METHODOLOGY

In conjunction to the advantages and limitations of both Hierarchical and Non Hierarchical approaches, a new mechanism called "Enhanced Hierarchical Agglomerative Clustering (EHAC)" which is Two Step Clustering process is introduced. In this method, we combined both Entropy based Mean Clustering proposed by V.V.Jaya Rama Krishniah and et al, and traditional Single –Link Hierarchical Agglomerative Clustering Approaches. On a Data Set, very first step we apply the Entropy based clustering and then in the second step, on resultant clusters of the step one, we applied the Hierarchical clustering. One of the major advantage with this approach is we can visualizes the non-global data,

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 3, September 2012

which is not possible with traditional non-hierarchical methods, interpretation of results is subject specific. And another advantage with proposed methodology is reduces the interpretation of complex hierarchies. The following Algorithm defines the mechanism involved in EHAC,

Algorithm

Input:
D = {d1, d2... dn} // set of *n* data items, Tree T = { }
Output: A set of *k* clusters.
Procedure
Phase 1

1. Fetch the each data element in the D and estimate the entropy of each data element.
2. Sort the data elements in descending order of entropies and call them as seeds.
3. Make the Candidate data set C such that no duplicates seed in C. and make one duplicate candidate set DC.

4. a) Set mean for each cluster CL_k as 0 and call it as Cluster Centre CC.

b) Assign a seed to every cluster CL_k. from the candidate set C.

5. Recompute the mean of each CL.

6. For each seed-point C*i* remain in C, find the closest centroid CC_j *and* assign C_i to cluster *j* .

7. a) Place the seed point C_i to the cluster CC_k such that the seed point distance is closer to the present nearest Distance.

b) Detach C_i from C

c) Repeat the step 5.

8. For each element in {D-DC} do the following step

a) Compare each CL_K seeds with the data seeds in {D-DC}.

b) Place the seeds in {D-DC} into the corresponding CL_k.

9. Repeat Step 6 to 8, until Candidate Set C becomes empty and convergence was made.

10. C={ C_1,C_2,C_3,C_4,....C_n}

Phase II

1. C'= {C_k/k∈ C} and T = { }, sequence node m =0 and Level L(0)=0
2. Do for every cluster C_K in C'

   a) Select any Cluster C_i in C' and find the least dissimilar pair of clusters in according to the distance between them d (C_i,C_k) = min d(C_i,C_k) and say best node

   b) Increment the sequence node m:=m+1 and create a new node HAC_m=C_iUC_k and set the level of cluster as L(m)=d(C_i,C_k)

   c) Calculate the Gap with neighboring nodes in the level L(m)

3. Repeat Step 2 until C' become empty (i.e. all the elements in one cluster).

The following figure shows the Block diagram form Enhanced HAC algorithm.

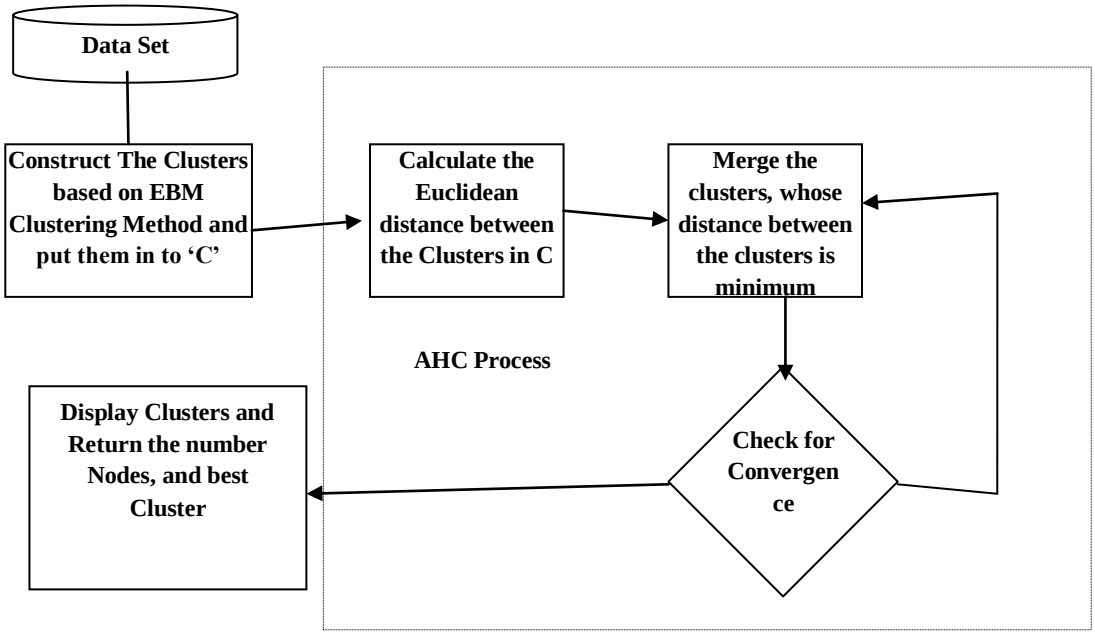

Fig 1: Block Diagram of EHAC

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 3, September 2012

III. RESULTS

This section provides the performance comparison of the conventional HAC and Enhanced HAC approaches. To make this evaluation, we used Hear Attack database extracted from UCI Machine Learning Database with 270 instances and 14 attributes. The following table shows the partial description of the database.

The performance of the proposed algorithm for the evaluation of Enhanced Hierarchical Algorithm (EHAC) and hierarchical clustering algorithms (HAC) is demonstrated on real data The important points to be checked are: (i) the ability of finding the correct number of clusters (ii) the quality of the results

Table 1: Heart Attack Database

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
70	1	4	130	322	0	2	109	0	2.4	2	3	3	2
67	0	3	115	564	0	2	160	0	1.6	2	0	7	1
57	1	2	124	261	0	0	141	0	0.3	1	0	7	2
64	1	4	128	263	0	0	105	1	0.2	2	1	7	1
74	0	2	120	269	0	2	121	1	0.2	1	1	3	1
65	1	4	120	177	0	0	140	0	0.4	1	0	7	1
56	1	3	130	256	1	2	142	1	0.6	2	1	6	2
59	1	4	110	239	0	2	142	1	1.2	2	1	7	2
60	1	4	140	293	0	2	170	0	1.2	2	2	7	2
63	0	4	150	407	0	2	154	0	4	2	3	7	2
59	1	4	135	234	0	0	161	0	0.5	2	0	7	1
53	1	4	142	226	0	2	111	1	0	1	0	7	1
44	1	3	140	235	0	2	180	0	0	1	0	3	1
61	1	1	134	234	0	0	145	0	2.6	2	2	3	2
57	0	4	128	303	0	2	159	0	0	1	1	3	1
71	0	4	112	149	0	0	125	0	1.6	2	0	3	1
46	1	4	140	311	0	0	120	1	1.8	2	2	7	2
53	1	4	140	203	1	2	155	1	3.1	3	0	7	2
64	1	1	110	211	0	2	144	1	1.8	2	0	3	1
40	1	1	140	199	0	0	178	1	1.4	1	0	7	1
67	1	4	120	229	0	2	129	1	2.6	2	2	7	2
48	1	2	130	245	0	2	180	0	0.2	2	0	3	1
43	1	4	115	303	0	0	181	0	1.2	2	0	3	1
47	1	4	112	204	0	0	143	0	0.1	1	0	3	1
54	0	2	132	288	1	2	159	1	0	1	1	3	1
48	0	3	130	275	0	0	139	0	0.2	1	0	3	1
46	0	4	138	243	0	2	152	1	0	2	0	3	1
51	0	3	120	295	0	2	157	0	0.6	1	0	3	1
58	1	3	112	230	0	2	165	0	2.5	2	1	7	2

The following sections describes about the different cluster results obtained from Agglomerative Hierarchical Approach and the Enhanced Hierarchical Clustering Approach, which is the two stage process combines both Entropy Based Mean Clustering and single-link Hierarchical Clustering approaches. To do this experimental study, we use Data mining Tool called Tangara for chart preparation and EBM Program, which is implemented in Java Language and MS-Excel for Data Representation.

A. Creating Clusters Using HAC

The following Tables 2 and 3 describes about the different features obtained by the true HAC Algorithm.

Table 2: Features of the HAC

Mechanism	No .of Instances	Feature				
		No Of Leaves	No Of Nodes	No of Clusters	Best Cluster	Gap
HAC	270	270	539	3	2	0.2891

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 3, September 2012

Table 3: Clusters in HAC

Cluster Number	Size	Center
cluster n°1	90	44.088889
cluster n°2	101	59.425743
cluster n°3	79	59.835443

The following figures 1 and 2 defines the clusters and corresponding Dendrogram of AHC

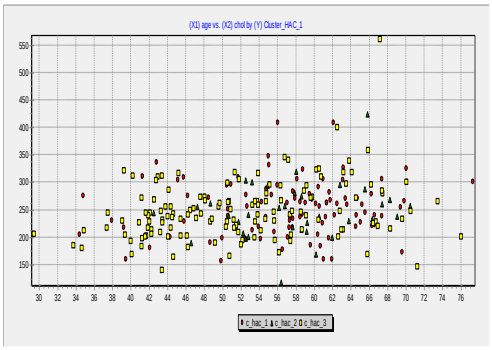

Fig 2: Clusters in the Data Set "Heart Attack" By Using AHC

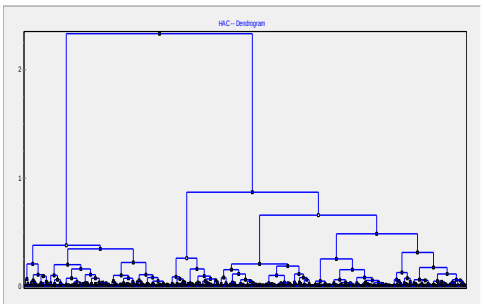

Fig: 3 Dendrogram for Data Set "Heart Attack" By Using AHC

B. ***Creating the Clusters by Using Entropy based Clustering Approach.***

The following Table 3 illustates about the clusters devloped by EBM Clustering, by using "age" attribute of the data set "Heart Attack".

Table 4: Clusters of Dataset Heart Attack by using EBM Clustering

Mechanism	No.of Instances	C=5 (No of samples)	C=10 (No of samples)	C=15 (No of samples)	C=20 (No of samples)
EBM Clustering	270	C1=125, C2=15, C3=84, C4=34, C5=12	C1=23, C2=15, C3=19, C4=12, C5=12, C6=42, C7=65, C8=11, C9=56, C10=15	C1=23, C2=15, C3=19, C4=12, C5=12, C6=35, C7=18, C8=11, C9=24, C10=15, C11=24, C12=9, C13=15, C14=30, C15=8	C1=22, C2=15, C3=12, C4=12, C5=12, C6=12, C7=11, C8=11, C9=24, C10=9, C11=24, C12=9, C13=15, C14=30, C15=8, C16=7, C17=7, C18=7, C19=16, C20=7

C. Clusters Using Enhanced HAC (EHAC)

The following tables 5 and 6 describes about different features and corresponding clusters of the Heart attack Dataset. In this phase, we applied the HAC Process on the clusters obtained by using EBM Clustering described in the Table 3

Table 5: Features of the EHAC Algorithm

Mechanism	No.of Instances	EBM Clusters	Feature				
			Number of Leaves	Number of Nodes	Number of Clusters	Best Cluster	Gap
Enhanced HAC	270	5	5	9	4	3	0.132
		10	10	19	3	3	0.132

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 3, September 2012

		15	15	29	3	1	0.132
		20	20	39	3	1	0.093

Table 6: Cluster and Size of the Clusters in EHAC Process

EHAC Size : 5				**EHAC Size : 10**			
Cluster	EBM Clusters	Size	Centers	Cluster	EBM Clusters	Size	Centers
C1	C1	125	47.08	C1	1,2,3,4,5,6,8,10	149	55.45
C2	C2,C5	27	58.00	C2	7	65	67.18
C3	C3	84	65.50	C3	9	56	39.81
C4	C4	34	54.58	-	-	-	-
EHAC Size : 15				**EHAC Size : 20**			
Cluster	EBM Clusters	Size	Centers	Cluster	EBM Clusters	Size	Centers
C1	1,2,3,4,5,6,7,8,9, 10,12,13,15	216	55.58	C1	1,2,3,4,5,6,7,8,9,10,12, 13,15,16,17,18, 19,20	216	55.66
C2	11	24	36.62	C2	11	24	37.71
C3	14	30	70.35	C3	14	30	69.28

The following figures 3 to Figure 10 illustrates about the clusters in the each EHAC size ranges with respect to EBM Clusters with size 5,10,15,20 and corresponding *Dendrogram.*

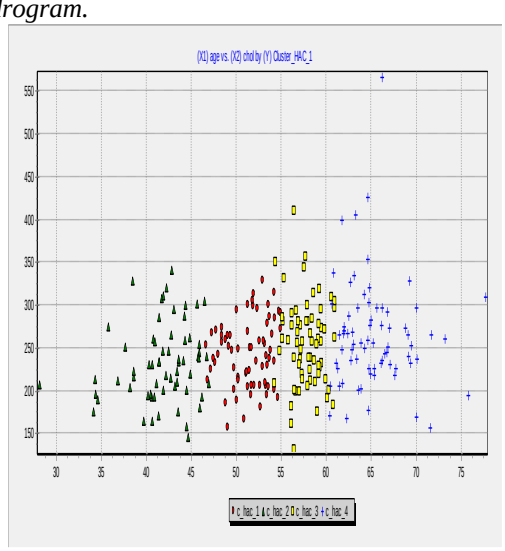

Fig 4: EHAC-5 Clusters

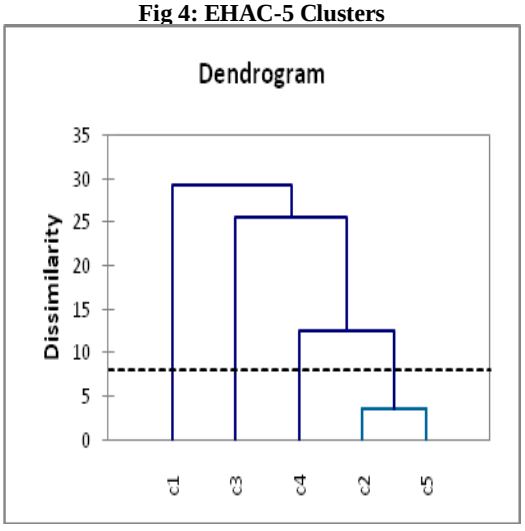

Fig 5: EHAC-5 Dendrogram

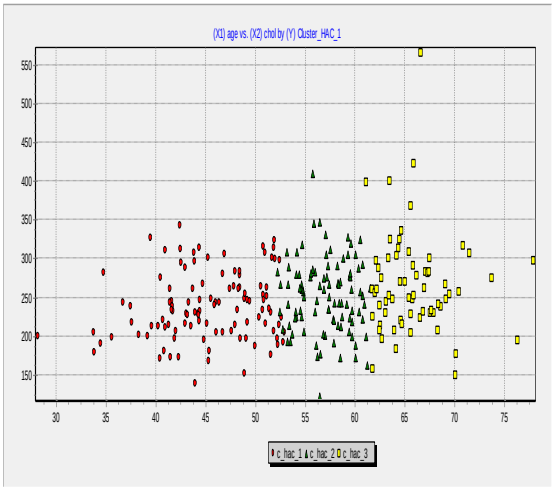

Fig6: EHAC: 10 Clusters

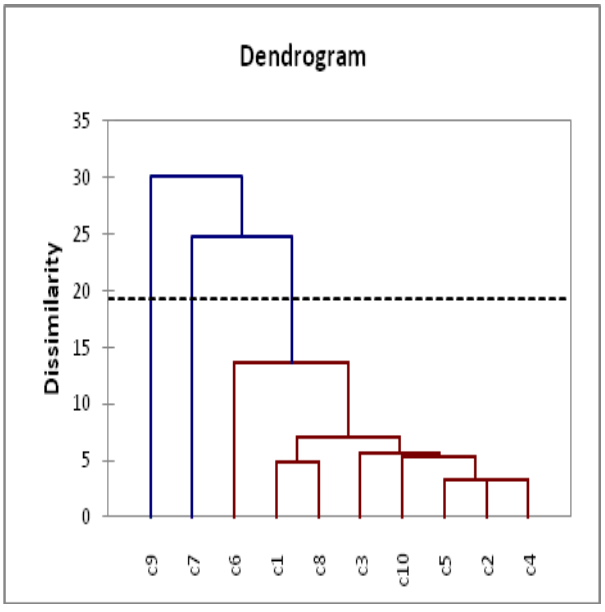

Fig 7: EHAC-10 *Dendrogram*

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 3, September 2012

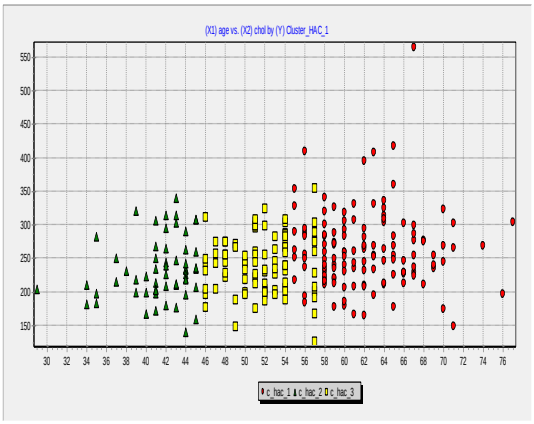

Fig 8: EHAC-15 Clusters

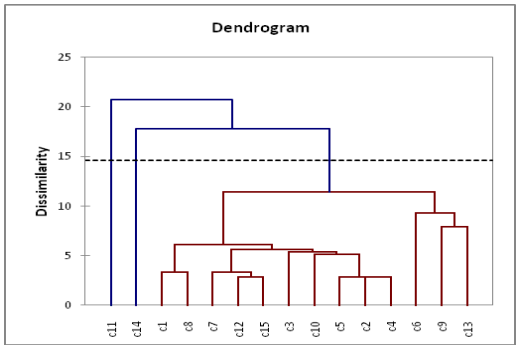

Fig 9: EHAC-15 *Dendrogram*

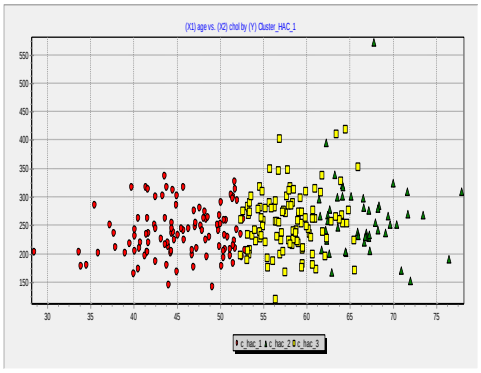

Fig10: EHAC: 20 Clusters

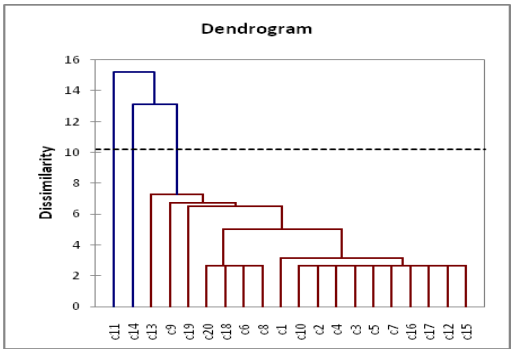

Fig11: EHAC20 Dendrogram

By Observing the Figure 2, we observe in AHC the number of levels in Dendrogram are very complex and around 539 nodes distributed over 10 levels. But from the figure2 5,7,19 and 11, we noticed only maximum 4 levels in the case of EHAC with EBM Size of 20 and the number of nodes were also correspondingly minimum as compared with AHC Algorithm. More over, with reference to the table 3 and table 4, we observe the centres of the clusters, we obtain in EHAC process were wide and accurate.

Table 7 Describes the Execution Time of the AHC and EHAC Approaches

AHC	EHAC Size: 5	EHAC Size: 10	EHAC Size: 15	EHAC Size: 20
4758	41	30	30	30

IV. CONCLUSION

The algorithm designed and developed in this paper for the evaluation of EHAC overcomes the complexity in traditional single-link Hierarchical clustering. Through the results produced from the proposed algorithm EHAC, one can easily understand different levels of observations on data analysis. One of the major advantages of this approach is, representation of Global clusters, which is not possible in the Non Hierarchical clustering approaches. The proposed algorithm can be optimized for feature selection, which reduces the dendrogram. In addition, we may increase performance of the EHAC by using Complete-Link intergroup similarity, which is one of the popular approaches in HAC.

REFERENCES

[1] Haifeng Zhao, Zijie Qi," Hierarchical Agglomerative Clustering With Ordering Constraints",

[2] K. Wag staff, C. Cardie, "Clustering With Instance-Level Constraints", Proceedings of the 17th International Conference on Machine Learning (ICML 2000), Stanford, Ca, Pp. 1103-1110. 2000.

[3] Ying Zhao, George Karypis, "Evaluation of Hierarchical Clustering Algorithms for Document Datasets".

[4] Hu Yang, Nicolino J. Pizzi,"Biomedical Data Classification Using Hierarchical Clustering" , IEEE,2004.

[5] Aleksander B. Demko, Nicolino J. Pizzi, Ray L. Somorjai," A System for the Analysis of Biomedical Data", IEEE Canadian Conference On Electrical and Computer Engineering, Pp. 1093.1098, 2002.

[6] Peter Bajcsy, Jiawei Han, Lei Liu, Jiong Yang, "Survey of Biodata Analysis from A Data Mining Perspective", *Springer link, 2005.*

[7] Roberto Avogadri, Matteo Brioschi, Francesca Ruffino, Fulvia Ferrazzi "An Algorithm to Assess the Reliability of Hierarchical Clusters In Gene Expression Data", Volume 5179, 2008, Pp 764-770.

[8] Bahadir Durak, "A Classification Algorithm Using Mahalanobis Distance Clustering Of Data with Applications on Biomedical Data Sets", 2011.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 3, September 2012

[9] Han, J., Kamber, M., "Data Mining: Concepts and Techniques", 2nd Edition, California: Morgan Kaufmann, 2006.

[10] K.P.Soman, Shyam Diwakar, V.Ajay, "Insight into Data Mining Theory and Practice", 2006.

[11] V.V.Jaya Ramakrishnaiah,Dr.K.Ramchand H Rao,Dr. R.Satya Prasad," Entropy Based Mean Clustering: A Enhanced Clustering Approach", Volume 1, No. 3, The International Journal Of Computer Science & Applications (TIJCSA), May 2012 Issn – 2278-1080.

AUTHOR BIOGRAPHY

V.V.Jaya Rama Krishaiah received Master's Degree in Computer Applications from Acharya Nagarjuna University, Guntur, India. Master of Philosophy from Vinayaka Missions University, Salem. He is currently working as Associate Professor in Department of Computer Science. ASN Degree College, which is affiliated to Acharya Nagarjuna University. He has 14 years of Teaching Experience. He is currently pursuing Ph.D., at Department of Computer Science and Engineering, Acharya Nagarjuna University, Guntur. His research area is Clustering in Databases. He has published several papers in National and International Journals.

D.V. Chandra Sekar received Master of Engineering with Computer Science and Engineering. He is currently working as Associate Professor, in the Department of Computer Science, TJPS College) PG Courses, Guntur, Which is affiliated to Acharya Nagarjuna University. Has 14 years of Teaching Experience and 1 Year Industry Experience. Has published 52 Papers in National and International Journals

Dr. K.Ramchand H Rao received Doctorate from Acharya Nagarjuna University, Masters Degree in Technology with Computer Science from Dr.M.G.R. University, Chennai, and Tamilnadu, India. He is currently working as Professor and Head of the Department, Department of Computer Science and Engineering, A.S.N. Women's Engineering College, Tenali, affiliated to JNTU Kakinada. He has 20 years of Teaching experience and2 years of Industry Experience at Morgan Stanly, USA, as Software Analyst. His research area is Software Engineering. He has published several papers in National and International Journals