International Journal of Advanced Engineering Research and Science (IJAERS) *[Vol-5, Issue-12, Dec- 2018]*
https://dx.doi.org/10.22161/ijaers.5.12.41 *ISSN: 2349-6495(P) | 2456-1908(O)*

Figuring out Extinct Values of Yeast Gene Microarray Expression (YGME) and Influencing Successive Time for Hierarchical Clustering Technique – An Improvement

Akey Sungheetha, Rajesh Sharma R

Department of Computer Science and Engineering, Adama Science and Technology University, Adama, Oromia.

Abstract**—** *The numerous missing value computation approaches for yeast data have been suggested in the literature. Throughout the past few years, investigators are keen on driving a lot of research effort on giving methodical assessments of the dissimilar computation procedures. The problem of controlling the missing values are designed with samples of tough microorganisms, such as yeast. Expensive strategies are present which has targeted to develop a varied collection of samples. They are regularly in effect for concurrently disturbing various small samples, but are greatly lesser effective for larger samples. The manufactured devices highlight interference rates after these minor samples having 5% of cells interrupted in 2 to 38 seconds range, frequently ignoring to indicate the organism interrupted or the small sample size. At the outset, maximum procedures continued to be evaluated by means of highlighting on the accuracy of the computation, using metrics such as the Correlation (uncentered), Correlation (centered), Absolute correlation (uncentered), Absolute correlation (centered), Spearman Rank correlation, Kendall's tau, Euclidean distance and City block distance. This proves the best clustering range. In the proposed approach running time is also computed for the various used methods using the same above mentioned metrics. On the other hand, it has turn out to be strong that the attainment of the accuracy and running time of the whole yeast gene data had a better assessment in further applied relations by way of hierarchical clustering approach. Accuracy and running time are sorted out for both large and small samples once after computing the missing values. Running times of the different clustering methods in a yeast dataset are existing in the work for the missing value rate of 4%. The hierarchical clustering was the fastest among the specified clustering methods (K-Means (gene) clustering technique, Self-Organized Mapping and Principle Component Analysis). However, the SOM was still about 10 times faster than k means. The running time of the original hierarchical method was about one third for that of its proposed version.*

Keywords**— Cluster, Yeast data, Hierarchical clustering, k means clustering, filtering data.**

I. INTRODUCTION

The greatest evidence result of small sample size, does not affect the quantification procedure. The whole yeast gene data are processed in the similar way from both small and large sample size. The missing value in the yeast gene data indications are visualized by reducing the dimensionality with hierarchical clustering approach. The objective of the research involves predicting the missing values and it is an essential step to determine missing values in microarray data as the whole dataset is necessary in several expression profile analysis in bioinformatics. Surely, any individual approach to confirm the investigation procedure of the microarray data with missing values is to repeat the computation, and evidently it is very costly and time consuming. Uniquely, one can be able to reflect, for instance, the capability clustering methods such as single linkage, complete linkage, average linkage and centroid linkage. These clustering methods of hierarchical clustering approach allows the dataset to preserve the important yeast gene data in the dataset, or its discriminative/predictive influence for classification/clustering determinations

The K-Means (gene) clustering technique, Self-Organized Mapping and Principle Component Analysis algorithms were clearly the slowest computation methods. The hierarchical clustering method made, on unusual case, assesses for missing values which were up to 4 times larger than the original values. This appears to put forward an inconsistency in the method's employment or process.

https://dx.doi.org/10.22161/ijaers.5.12.41 *ISSN: 2349-6495(P) | 2456-1908(O)*

Integrative Missing Value Assessment through hierarchical clustering is the initial technique to include data of microarray datasets to improve missing data computation [1]. Though, it is hard to discover data in the datasets and even further demanding to discover a set of genes often indicate expression resemblance to the target gene over numerous genes. In the meantime, centroid linkage, single linkage, complete linkage and average linkage are the foremost algorithm that exploits the useful similarities fixed in the yeast microarray data along with the expression similarities to enable the neighbor gene selection [2]. It outperformed k means, at high missing percentages, owing to the control of the amount and accuracy of the gene utilities interpreted in yeast data, Self-Organized Mapping and Principle Component Analysis algorithms miscarried to improve the time consumption in the computation process.

To the understanding, first study has inspected the consequence of missing values and their computation on the maintenance of clustering results. Other studies determined missing values on K-Means (gene) clustering technique, Self-Organized Mapping and Principle Component Analysis computation method did not deliberate genetic analysis on the clustering results; their core outcomes were that even a small amount of missing values may intensely drop the steadiness of K-Means (gene) clustering technique, Self-Organized Mapping and Principle Component Analysis computation and hierarchical clustering algorithms evidently recover this steadiness [3]. Hence the outcomes are in worthy with these conclusions.

The three steps to retrieve data are Loading, Filtering and Adjusting Data in clustering. Information in the form of dataset are loaded and processed as a Cluster. The four clustering methods such as centroid linkage, single linkage, complete linkage and average linkage are provided for adjusting and filtering the data that has been loaded. These methods gain access to Filter Data and Adjust Data. Filtering data permits to get rid of yeast gene expression datas that ensure not satisfy certain desired conditions. Adjusting data leads to perform conditional operations. The primary choice made essential is how similarity between yeast gene expression data expression data is to be well-defined. There are several methods to compute exactly how comparable two series of records are. Cluster provides eight options namely Correlation (uncentered), Correlation (centered), Absolute correlation (uncentered), Absolute correlation (centered), Spearman Rank correlation, Kendall's tau, Euclidean distance and City block distance.

II. RELATED WORKS

There are several computation techniques have been proposed since 1963, such as hierarchical grouping, hierarchical clustering, and since 2009 such as K-Means (gene) clustering technique, Self-Organized Mapping and Principle Component Analysis. [4, 5, 6, 7, 8, 9]. The most commonly used technique among these is the hierarchical clustering. However all of the methods of hierarchical such as centroid linkage, single linkage, complete linkage and average linkage measures are merely recognized on the yeast gene expression datasets themselves and employ nothing of the external microarray datasets or genetic associated data. Here numerous modest methods are present to determine the missing values, e.g. eliminating the genes with missing values from supplementary study, substituting missing values by zeros, or satisfying the missing values with the row or column means/medians present [10, 11, 12]. These methods are not ideal as they did not deliberate the relationship of the data, which stimulated the progress of further refined missing value ways that strained to exploit the data associations by means of the data present in the entire dataset [13].

As per data given in Table 1, missing value is a common difficulty that has to be addressed even for further modern educations [14, 15]. Likewise, here exists several genes with high missing percentages. In this circumstance, for genes with numerous missing values, little values are persisted to conclude in what way the gene is associated with other genes in the dataset, which leads to less accurate assessments. It is well known that gene expressions in cells are concertedly measured by similarity factors and information encoded in the nuclear and mitochondrial genomes of the yeast [16]. The major iterating unit of mitochondrial genomes, which consists of approximately 1000's of microorganisms around Genome Database [17]. For instance as mentioned in [18, 19, 20], mitochondrial genomes might modify the structure. Thus, the similarity factor is greatly measured by the mitochondrial genomes states in mitochondrial. Nevertheless, definite objective existed to examine the consequence of missing values on the hierarchical clustering algorithms, such as centroid linkage, single linkage, complete linkage and average linkage, and to discover whether new progressive computation methods, such as SOM, will be able to offer improved clustering results than the old-style k means method. The outcomes recommend that hierarchical clustering runs fast, robust and accurate outcomes, particularly when the missing value rate is lower than 4%. None of the computation methods might sensible and correct for the stimulus of missing values above this 4% threshold. In these circumstances, one must think through in eliminating the genes with many missing values or iterating the tests if

https://dx.doi.org/10.22161/ijaers.5.12.41 *ISSN: 2349-6495(P) | 2456-1908(O)*

likely. As prominent before, clustering related to datasets are naturally regularized therefore a data value near to zero shows the nonexistence of any relations in the midst of a pair of genes. Thus a simple key to the problem of missing values is to substitute those items with zeros. However this might give the impression to be a hierarchical cluster methodology, it has some validation: the probability is that maximum genes do not work together, and hence their relations score is probably to be close to zero. Likewise it is perceived that the mean /median of the non-missing entries in the datasets defined before is almost zero. This method helps as a starting point for investigational assessments.

Loading, Filtering and Adjusting Data: A machine learning system is established for deciding gene functions from assorted source of data sets using hierarchical clustering. Through a prearrangement, in the Group of input data tables rows signify genes and columns denote samples or interpreted values known as yeast data microarray hybridization. On performing the three steps to retrieve data namely Loading, Filtering and Adjusting Data in clustering, a small size Cluster input data resembles as in Table 1 [21].

Loading data: The YORF field contains an alpha-numeric value. It is forecasted in Tree View to state how the rows are connected. The left over chambers in the table contain data for the suitable gene and sample. The readings are observed as data for instance 1 at 0 min for YAL001w and missing value for gene YAL001C at 2 hours was 5.8. Omitted data are tolerable and are nominated by blank cells. In order to identify the missing value, the operation "Present % >= X" is enabled.

Table. I YORF-Yeast open reading frame

S. No	(YORF)	O min	30 min	1 hour	2 hour	4 hour
1	YAL001w	1	1.3	2.4	5.8	2.4
2	YAL002w	0.9	0.8	0.7	0.5	0.2
3	YAL003W	0.8	2.1	4.2	10.1	10.1
4	YAL005C	1.1	1.3	0.8		0.4
5	YAL010C	1.2	1	1.1	4.5	8.3

The large size sample data file similar to small size sample data file as given in Table I comprises yeast gene expression data defined in Eisen et al. Move this data to testing and training in addition to loading the Cluster bunch. Each Cluster bunch resolved will provide information roughly about the loaded data file. Once loaded, the listed, used and calculated measures such as Correlation (uncentered), Correlation (centered), Absolute correlation (uncentered), Absolute correlation (centered), Spearman Rank correlation, Kendall's tau, Euclidean distance and City block distance are used as the testing and training statistics for different cluster analytical methods. Grouping is a significant tool for exploring such Cluster bunch of microarray information, usual properties of which are its intrinsic ambiguity, noise and fuzziness [22, 23, 24, 25, 26, 27, 28]. The columns and rows in the dataset are elective. Hence the Tree View practices to use the ID in YORF column by the means of labelling for each individual gene and YORF column permits to identify a label for each individual gene that is isolated after the ID is specified in the YORF column. The 31 rows and 79 columns will be labelled well ahead in the dataset for loading purpose. The Filter Data permits to take out the genes that do not take part definitely sought after setting the properties of dataset. The properties such as enable and disable options are used to load, apply filter and accept filter as shown in Table II.

Filter data: The filtering of data is the process of eliminating genes that abstain in certain preferred properties which is described in Table II. Also the presently accessible properties that can be capable to be used to filter data are existing [28]. These stay impartially understandable. As soon as filter are implemented, the filters are not instantly used in the dataset. Primarily the filter implementation expresses exactly how many genes would have been accepted by the filter. If accepted, genes passes through the filter, or else certainly no modifications are made.

Table. II Eliminate genes lacking desired properties from dataset of 31 rows and 79 columns

S.No	Limitation	Status
1	Present % >= 80=A	Enabled
2	SD(Gene vector)>2.0=A	Disabled
3	At least 1 observation with abs(Val)>=2.0=B	Disabled
4	High Value-Low Value>=2.0=A	Disabled
5	Apply filter	21 passed out of 31
6	Accept Filter	Enabled

- Step 1 eliminates the entire genes that have missing numerical information in larger than (100 − A) percentage of the columns.

https://dx.doi.org/10.22161/ijaers.5.12.41 *ISSN: 2349-6495(P) | 2456-1908(O)*

- Step 2 eliminates the entire genes that have normal abnormalities of detected numerical information lesser than A.
- Step 3 eliminates the entire genes that do not have minimum of A interpretations analysed through total numerical information larger than B.
- Step 4 eliminates the entire genes whose higher value subtracts the low value that are less than A.

The genes are passing the filter when applying and accepting the filter. In order to filter data, the default value is set to read the result. Hence they are kept NIL as given in Equation 1 and Equation 2.

Apply filter = NIL .....Eq 1

Accept filter = NIL .....Eq 2

The default values are presented in Table 3 for passing the genes through the filter.

Table. III Assign default values to filter genes lacking desired properties from dataset of 31 rows and 79 columns

S.No	Option	Entry	Value
1	Disabled	% present>=	80
2	Disabled	SD (Gene vector)	2.0
3	Disabled	At Least	1
4	N/A	Observation with abs (val)>=	2.0
5	Disabled	High Value-Low Value>=	2.0

There are six conditions to pass the genes through the filter. They are illustrated as follows:

Condition 1: After applying filter operation for the given dataset with an assigned default value as given in Table III, then the numerical information in the entire 31 rows passes out of 31 rows without any missing information. It is found that there are no missing values. This is proved by identifying the result through the gene cluster tool. Hence the result is presented in Table IV.

Table. IV Identifying genes lacking desired properties from dataset of 31 rows and 79 columns > 100-80

S.No	Option	Entry	Value
1.	Enabled	% present>=	80

Condition 2: Next, if the genes have %present >=80, then the result shows that it has no missing information and also filtering task is not further necessary while passing the genes.

Condition 3: This condition where, if the abnormality of the Standard deviation, SD (gene vector) is enabled, none of the numerical information passes out of 31 rows. Then all the detected numerical information less than 2.0 (SD) are removed.

Condition 4: The genes are passed for at least 1 observation by means of total absolute value as given in the Equation 3, where the abs(Val) is larger than 20, which allows 3 rows to pass out of 31 rows.

abs(Val)>20 .....Eq 3

Condition 5: The filtered gene for the High value is subtracted from Low Value as given in Equation 4.

High Value-Low Value > = 2.0 .....Eq 4

This condition also passes 3 rows out of 31 rows similar to condition 4.

Condition 6: If the filtered genes have high value as given in Equation 5 then the filter passes 21 rows passed out of 31 rows.

High Value>=20 .....Eq 5

Finally, the filter process is accepted for condition 3, 4, 5 and 6 in order to accept filtering rows further.

Adjust Data-Units mean: There are five number of tasks used to adjust the information and the tasks are performed by modifying the original information. The information is adjusted interms of log transform data, center gene-mean, center arrays-mean, normalizing gene and normalizing arrays subsequently the middle gene and middle array imperative process has its median for an assessment to fine-tune information.

III. PROPOSED STUDY ON CLUSTERING FOR SMALL SAMPLE SET -HIERARCHICAL (GENE) CLUSTERING

The procedures for establishing hierarchical clusters are of commonly private subgroups (genes and arrays). An individual of private subgroups which has members that are extremely alike with an esteem are used to identify features integrating nearest neighbour searching algorithm. These weights are determined in addition to grouping [29,30]. Then the cutoff value (0.1) and the exponent value (1) are set as a default value and the similarity metric measure, correlation uncentered is chosen for determining the weights. The correlation (uncentered) metric is the one that rely on centroid linkage where a vector is assigned to compute the distance. The distances are computed with the centroid linkage method that will cluster and generate the cluster bunch. Firstly, the gene tree file (.gtr) is generated with node and gene value with its exponent. Secondly, an array tree (.atr) disk image (a copy of 8 bit formatted disk) file is generated with node and its array value with the same exponent 1. Thirdly, a coral draw text editor image template (.cdt) is generated with the E weight (exponent weight) of G weight (Gene Weight). The similar performance process of generating files for the centroid

https://dx.doi.org/10.22161/ijaers.5.12.41 ISSN: 2349-6495(P) | 2456-1908(O)

linkage method in hierarchical clustering is followed to single linkage method, complete linkage method and average linkage method. For instance, the centroid linkage method involves two node and two gene value for generated gene tree as shown in Table V (as sample1).

Table.V Node gene Sample 1

Node	Gene Matrix		Range
Node 1x	Gene 0x	Gene 1x	-0.527353
Node 2x	Gene 1x	Gene 2x	-0.94495

The interference for the single linkage method is derived as given in Table VI (ie.,sample2).

Table.VI Node gene Sample 2

Node	Gene Matrix		Range
Node 1x	Gene 0x	Gene 1x	-0.527353
Node 2x	Gene 1x	Gene 2x	-0.611316

The rest of the files are similarly generated for centroid linkage and single linkage method. The complete linkage method differs in value from others. It generates the value as given in Table VII (ie., sample 3).

Table..VII Node gene Sample 3

Node	Gene Matrix		Range
Node 1x	Gene 1x	Gene 0x	-0.527353
Node 2x	Gene 2x	Gene 1x	-0.819574

For average linkage method, gene tree file is generated as given in Table VIII (ie., sample 4) showing one different value for the second node similar to other two methods.

Table.VIII Node gene Sample 4

Node	Gene Matrix		Range
Node 1x	Gene 0x	Gene 1x	-0.527353
Node 2x	Gene 1x	Gene 2x	-0.715445

After performing hierarchical clustering, k-means clustering is chosen for evaluation. The similar dataset of Eigen which is fed for hierarchical clustering is used in k-means clustering.

K-Means (gene) clustering technique: The genes and arrays of the dataset are analysed using the k-mean clustering algorithm. Both genes and arrays have 10 numbers of cluster k and 100 numbers of runs each where the k-means and k-medians are determined. On execution of k-means with the Euclidean distance similarity metric for both gene and array, it is found that clusters are available more in number than the genes. Then the entire dataset is passed without any gene filter irrespective of number of observations or absolute value specification. Also, the data is adjusted and it is independent of hierarchical technique. After execution, the cluster k generates a cluster gene file (.kgg) where gene groups 10 clusters and the data in open reading frame (ORF) is a .kgg file and .kag file. It groups the gene into 10 groups and Cluster, k for 10 gene and 10 array are listed with gene weight and experiment weight.

Self-Organized Mapping and Principle Component Analysis: After the execution of k-means clustering technique, the same Eisen dataset is tested in Self Organized Mapping (S0M) and Principle Component Analysis (PCM). The SOM organizes the genes and arrays similar to k-means clustering. The X dimension and Y dimension are assigned for the genes and arrays (as 3). The number of iterations for genes by default is 1, 00,000 and arrays is 20, 000 respectively. The initial tau is set to 0.02 by default and the outcome of both the genes and arrays of SOM are similar. The similarity metric here is the Euclidean distance and the three files generated of which GNF file shows the gene vectors and ANF file shows the array vectors. The gene/array file together shows the gene weight and experiment weight of the vectors. The mean values are not presented in the self-organized maps [31]. So the clustering technique of principle component analysis (PCA) is applied for Genes & Arrays to calculate the mean. PCA execution results in generating the principle component of array and gene. The gene and array are coordinating in two ways. The array co-ordinate is showing Eigen value of experiment weight and gene co-ordinate showing gene weight. All the clustering technique such as hierarchical, k-mean, self-organized mapping and PCA have adjusted the data to the mean. When adjusting data to median the result on filter data is as shown below. Hence the tata must be filtered before adjusting process.

Filter data: Filtering the data with mean is similar to the process of filtering the data with median.

Adjusting data with median for Atleast 1 observation with abs(val)>=2.0

The difference discovered in filtering data with mean and median shows that when adjusting mean first and then filtering, shows no rows have passed out of 31 rows. Adjusting median first and then filtering also shows no rows have passed out of 31 rows. When filtering gene for at least 1 observation with abs(val)>=2.0 shows 3 rows passing out of 31 rows. The filter is being accepted to perform clustering after the rows are passed. Adjusting the data for the center gene and center array to mean and median respectively and vice versa filter no rows have

https://dx.doi.org/10.22161/ijaers.5.12.41 *ISSN: 2349-6495(P) | 2456-1908(O)*

passed out of 31 rows. Adjusting data with median is similar to adjusting data with mean in log transform data and normalizing gene or arrays for center genes and center arrays respectively.

IV. PROPOSED STUDY ON CLUSTERING FOR HIERARCHICAL (GENE) CLUSTERING TECHNIQUE - LARGE SAMPLE SET

The various similarity metric performances are measured. They are: Correlation (uncentered), Correlation (centered), Absolute correlation (uncentered), Absolute correlation (centered), Spearman Rank correlation, Kendall's tau, Euclidean distance and City block distance.

https://dx.doi.org/10.22161/ijaers.5.12.41 *ISSN: 2349-6495(P) | 2456-1908(O)*

Table. IX Comparison between clustering methods

Clustering method	Gene/array similarity metric	31 rows node/gene		31rows node/array		2467 rows node/gene		2467 rows node/array	
Centroid linkage	Correlation uncentered	0.642641	0.16757	0.90082	0.668934	0.988387	0.354391	0.929455	0.075474
Single linkage		0.642641	0.336574	0.90082	0.722635	0.988387	0.414903	0.929455	0.288938
Complete linkage		0.642641	-0.34663	0.90082	-0.805213	0.988387	-0.883172	0.929455	-0.489157
Average linkage		0.642641	0.100977	0.90082	-0.110935	0.988387	-0.28906	0.929455	0.0223
Centroid linkage	Correlation centered	0.640981	0.123294	0.896823	-0.541497	0.989404	-0.606245	0.926293	-0.141204
Single linkage		0.640981	0.287747	0.896823	0.418646	0.989404	0.961167	0.926293	0.287638
Complete linkage		0.640981	-0.335755	0.896823	-0.750119	0.989404	-0.89763	0.926293	-0.520852
Average linkage		0.640981	0.090961	0.896823	-0.082129	0.989404	-0.068484	0.926293	-0.018541
Centroid linkage	Absolute correlation uncentered	0.642641	0.167570	0.900820	0.063715	0.988387	0.094143	0.929455	0.054159
Single linkage		0.642641	0.336574	0.900820	0.444248	0.988387	0.414903	0.929455	0.332774
Complete linkage		0.642641	0.000931	0.900820	0.000000	0.988387	0.000000	0.929455	0.000056
Average linkage		0.642641	0.130989	0.900820	0.158227	0.988387	0.114757	0.929455	0.092952
Centroid linkage	Absolute correlation centered	0.640981	0.123294	0.896823	0.018289	0.989404	0.013264	0.926293	0.071195
Single linkage		0.640981	0.293646	0.896823	0.418646	0.989404	0.404155	0.926293	0.335699
Complete linkage		0.640981	0.001184	0.896823	0.000083	0.989404	0.000000	0.926293	0.000074
Average linkage		0.640981	0.117903	0.896823	0.152002	0.989404	0.126558	0.926293	0.087962
Centroid linkage	Spearman rank correlation	0.693216	-0.049660	0.910012	-0.274194	0.973099	-0.001144	0.906171	-0.126924
Single linkage		0.693216	0.283253	0.910012	0.337878	0.973099	0.412512	0.906171	0.265874
Complete linkage		0.693216	-0.414645	0.910012	-0.691423	0.973099	-0.818796	0.906171	-0.477460
Average linkage		0.693216	0.064168	0.910012	-0.051662	0.973099	-0.024957	0.906171	-0.022292
Centroid linkage	Kendall's tau	0.508900	-0.056484	0.746514	-0.135484	0.885758	0.011360	0.749915	-0.085986
Single linkage		0.508900	0.195261	0.746514	0.246734	0.885758	0.296595	0.749915	0.183322
Complete linkage		0.508900	-0.267782	0.746514	-0.510871	0.885758	-0.636636	0.749915	-0.340330
Average linkage		0.508900	0.044218	0.746514	-0.037265	0.885758	-0.002423	0.749915	-0.015095
Centroid linkage	Euclidean distance	0.928197	0.000000	0.954213	0.000000	0.995196	0.000000	0.928997	0.000000
Single linkage		0.914380	0.000000	0.924451	0.000000	0.991290	0.000000	0.894987	0.000000
Complete linkage		0.950895	0.000000	0.981846	0.000000	0.998144	0.000000	0.978606	0.000000
Average linkage		0.936194	0.000000	0.964699	0.000000	0.995290	0.000000	0.956165	0.000000
Centroid linkage	City block distance	0.732687	0.000000	0.775965	0.000000	0.928988	0.000000	0.737505	0.000000
Single linkage		0.712875	0.000000	0.690699	0.000000	0.909338	0.000000	0.675318	0.000000
Complete linkage		0.774877	0.000000	0.867530	0.000000	0.960542	0.000000	0.854260	0.000000
Average linkage		0.748284	0.000000	0.795720	0.000000	0.932525	0.000000	0.791840	0.000000

Table IX gives a comparison of similarity measure performance on different clustering methods. Also it helps in identifying the missing values of yeast which leads to determine the time complexity.

V. RESULTS AND DISCUSSION

Clustering gene and array with hierarchical technique sorts with similarity metric correlation (uncentered) for centroid linkage clustering method. It results in sorting from 0.642641 to 0.167570 (node/gene) for instance.

Table. X The codes for the methods

Method	code
Hierarchical	H
Gene	G
Clustering	C
Gene Array	GA
Correlation (uncentered)	CU
Correlation (centered)	CC
Absolute correlation (uncentered)	ACU

https://dx.doi.org/10.22161/ijaers.5.12.41 *ISSN: 2349-6495(P) | 2456-1908(O)*

Absolute correlation (centered)	ACC
Spearman Rank correlation	SRC
Kendall's tau	KT
Euclidean distance	ED
City block distance	CBD
Centroid Linkage	CEL
Single Linkage	SL
Complete Linkage	COL
Average Linkage	AL
Cluster	C
Cluster Weights	CW

For single linkage the corresponding node/gene, node/array and the weights are presented in the tabulation for the method (H_G_C_CU_SL).

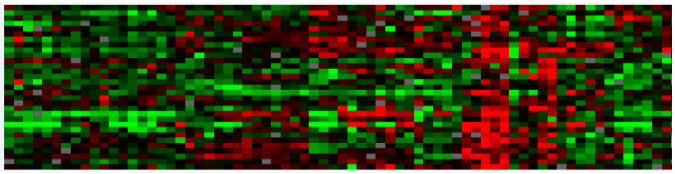

Fig 1. Gene tree view 31rows 79columns

For complete linkage and average linkage, H_G_C_CU_COL and H_G_C_CU_AL, the same evaluation is done as in centroid and single linkage. All these methods are tested for all the other similarity metrics and the performance is updated in Table V. For correlation centered, the corresponding procedure code H_GA_C_CC_CEL,H_GA_C_CC_SL,H_GA_C_CC_C OL and H_GA_C_CC_AL are used. The range of node/gene for H_GA_C_CU_CEL and H_GA_C_ACU_CEL are the same. The initial value of node/array range for H_GA_C_CU and H_GA_C_ACU are same in all four methods (centroid, single, complete and average).

Table. XI 31rows and 2467 rows (79 columns) – cluster range

Cluster method	**Range**	**31 rows node/gene**	**31 rows node/array**	**2467 rows node/ gene**	**2467 rows node/ array**
	SR	0.642641	0.9008	0.988387	0.929455
CEL	ER	0.16757	0.6689	0.354391	0.075474
SL	ER	0.336574	0.7226	0.414903	0.288938
CL	ER	-0.34663	-0.805	-0.88317	-0.48916
AL	ER	0.100977	-0.111	-0.2890	0.0223
				6	

The small scale information involve the observations for only 31rows 79columns. On increasing the size to 2467 rows 79 columns as given in Table XI, clustering performance is maintained in an effective way such that the Euclidian and city block distance measure with large dataset shows better outcome when compared to other similarity measures [32-34]. The time taken to cluster data with the similarity measures ACC, SRC and KT are determined. Also the ACU, ACC, ED and CBD time computation is calculated for the gene/array cluster bunch that involve the weight of cutoff=0.1 and exponent=1 for gene and arrays. Only few similarities and variations are noted in case of CU on comparing two values C and CW, the starting value range for the cluster is nearer to cluster weights for CEL.

Table XII. 31rows and 2467 rows (79 columns) – execution time

Cluster Metrics	**Time (sec)**			
	31 rows node/ gene	**31 rows node/ array**	**2467 Rows node/ gene**	**2467 rows node/ array**
Correlation (uncentered)	38	35	31	34
Correlation (centered)	32	34	30	33
Absolute correlation (uncentered)	30	28	29	27
Absolute correlation (centered)	26	28	28	26
Spearman Rank correlation	22	25	28	24
Kendall's tau	21	23	22	22
Euclidean distance	10	2	4	6
City block distance	7	4	5	2

On comparing the time taken to execute clustering using ED and CBD measure, it takes very less duration to process the data as given in Table XII.

For comparing these techniques used in this work, a statistical test has been conducted. Z-test for testing equality of variance between the similarity measures has been used to test the hypothesis of equality of two population variances shows 6.25 for Correlation (uncentered) and 8.25 for Correlation (centered) and no

https://dx.doi.org/10.22161/ijaers.5.12.41 *ISSN: 2349-6495(P) | 2456-1908(O)*

variances for Absolute correlation (uncentered), Absolute correlation (centered), Spearman Rank correlation, Kendall's tau, Euclidean distance and City block distance when the sample size of each sample is 30 or larger.

VI. CONCLUSION

Similar to CU, the SR for CC, ACU, ACC, SRC and KT similarity measures are same. The ER differs for CC, ACU, ACC, SRC and KT. In case of ED and CBD, the SR for cluster methods is different and ER is same. The time taken for KT alone takes more time to generate the output. The gene tree view for 31rows 79columns with x and y pixels, mask<0 and corr select cutoff=0.8 are shown in Figure 1. The colour indications are green-negative, black-zero, red-positive and gray missing. The gene tree view for 2467rows and 79columns have reduced missing values. Hence the data mining methods are studied and compared for measuring clustering performance for various methods.

The future progress can be tested with same small and large sample yeast gene data for self-organized mapping and principle component analysis. It uses the similar process that has been used in hierarchical and k means clustering. Also the performance time can be reduced.

REFERENCES

[1] Rajesh Sharma R, Akey Sungheetha, Dual Tree Complex Wavelet Transform, Probabilistic Neural Network and Fuzzy Clustering based on Medical Images Classification – A Study, International Journal of Advanced Engineering, Management and Science, vol. 4, no. 12, pp. 793-799 (2018),

[2] Tuikkala J, Elo L, Nevalainen OS, Aittokallio T: Improving missing value estimation in microarray data with gene ontology. Bioinformatics 2006, 22(5):566–572.

[3] Sharma, R. Rajesh, and P. Marikkannu. "Hybrid RGSA and support vector machine framework for three-dimensional magnetic resonance brain tumor classification." ScientificWorldJournal 2015 (2015): 184350.

[4] Sungheetha, Akey, and J. Suganthi. "An efficient clustering-classification method in an information gain NRGA-KNN algorithm for feature election of micro array data." Life Sci J 10.Suppl 7 (2013): 691-700.

[5] Sharma, Rajesh, and Akey Sungheetha. "Segmentation and classification techniques of medical images using innovated hybridized techniques—a study." Intelligent Systems and Control (ISCO), 2017 11th International Conference on. IEEE, 2017.

[6] Sungheetha, Akey, and R. Rajesh Sharma. "Extreme Learning Machine and Fuzzy K-Nearest Neighbour Based Hybrid Gene Selection Technique for Cancer Classification." Journal of Medical Imaging and Health Informatics 6.7 (2016): 1652-1656.

[7] Beaula, A. Rajesh Sharma R., et al. "Comparative study of distinctive image classification techniques." Intelligent Systems and Control (ISCO), 2016 10th International Conference on. IEEE, 2016.

[8] J. Suganthi Akey Sungheetha, "Energy Saving Optimized Polymorphic Hybrid Multicast Routing Protocol." International Review on Computers and Softwares, Vol.8, No.6, pp. 1367 – 1373.

[9] Sungheetha, A., Mssujitha, R., Arthi, V., Sharma, R.R. 2017, Data analysis of multiobjective density based spatial clustering schemes in gene selection process for cancer diagnosis, Proceedings of 2017 4th International Conference on Electronics and Communication Systems, ICECS 2017.

[10] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JG, Sabet H, Tran T, Yu X, Powell JI, Yang LM, Marti GE, Moore T, Hudson J, Lu LS, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000, 403(6769):503–611.

[11] Schafer JL, Graham JW: Missing data: our view of the state of the art. Psychol Methods 2002, 2(7):147–177.

[12] Little RJA, Rubin DB: Statistical analysis with missing data. New York: John. Wiley & Sons; 1987.

[13] Meneghini MD, Wu M, Madhani HD: Conserved Histone Variant H2A.Z Protects Euchromatin from the Ectopic Spread of Silent Heterochromatin. Cell 2003, 112: 725–736.

[14] Kobor MS, Venkatasubrahmanyam S, Meneghini MD, Gin JW, Jennings JL, Link AJ, Madhani HD, Rine J: A Protein Complex Containing the Conserved Swi2/Snf2-Related ATPase Swr1p Deposits Histone Variant H2A.Z into Euchromatin. PLoS Biol 2004.

[15] Yuan GC, Ma P, Zhong WX, Liu JS: Statistical assessment of the global regulatory role of histone acetylation in Saccharomyces cerevisiae. Genome Biol 2006, 7: 8.

[16] Yuan GCLY, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ: Genome-scale identification of nucleosome positions in S. cerevisiae. Science 2005, 309: 626–630.

https://dx.doi.org/10.22161/ijaers.5.12.41 *ISSN: 2349-6495(P) | 2456-1908(O)*

[17] Schubeler D, MacAlpine DM, Scalzo D, Wirbelauer C, Kooperberg C, van Leeuwen F, Gottschling DE, O'Neill LP, Turner BM, Delrow J, Bell SP, Groudine M: The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. Genes Dev 2004, 18(11):1263–1271.

[18] Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, Lee TI, Bell GW, Walker K, Rolfe PA, Herbolsheimer E, Zeitlinger J, Lewitter F, Gifford DK, Young RA: Genome-wide map of nucleosorne acetylation and methylation in yeast. Cell 2005, 122(4):517–527.

[19] Rando OJ: Global patterns of histone modifications. Curr Opin Genet Dev 2007, 17: 94–99.

[20] Rajesh Sharma R, P. Marikkannu, Akey Sungheetha, "Three-Dimensional MRI Brain Tumor Classification using Hybrid Ant Colony Optimization and Gray Wolf Optimizer." International Journal of Biomedical Engineering and Technology, vol. 29, no. 1, pp. 34-45 (2019).

[21] Bandyopadhyay S, Mukhopadhyay A, Maulik U, (2007), An improved algorithm for clustering gene expression data. Bioinformatics, vl. 23(21), pp. 2859-2865.

[22] Rao A, (2002), A clustering algorithm for gene expression data using wavelet packet decomposition, Systems and Computers, Conference Record of the Thirty-Sixth Asilomar Conference on IEEE,Vol. 1, pp. 316-319.

[23] Tseng G,(2004), A comparative review of gene clustering in expression profile,Automation, Robotics and Vision Conference, ICARCV 8th IEEE,Vol. 2, pp. 1320-1324.

[24] Chow C, K Zhu, H Lacy, J Lingen, M W, Kuo,(2009), A cooperative feature gene extraction algorithm that combines classification and clustering, In Bioinformatics and Biomedicine Workshop, BIBMW International Conference on IEEE,pp. 197-202.

[25] Dutta, Dipankar, Pranab Dutta, and Jaya Sil. "Data clustering with mixed features by multi objective genetic algorithm." Hybrid Intelligent Systems (HIS), 2012 12th International Conference on. IEEE, 2012.

[26] Choudhury N, Sarmah R, & Sarma S, (2012), A modified QT-clustering algorithm over Gene Expression data, In Recent Advances in Information Technology (RAIT),1st International Conference on IEEE,pp. 542-547.

[27] Eisen M B, Spellman P T, Brown P O, & Botstein D , (1998), Cluster analysis and display of genome-wide expression patterns, Proceedings of the National Academy of Sciences, 95(25), pp. 14863-14868.

[28] Sharma, Rajesh, P. S. Renisha, and Akey Sungheetha. 2016 "Comparative Study on Medical Image Classification Techniques." International Journal of Advanced Engineering, Management and Science 2.11.

[29] Sharma, Rajesh, et al. 2016 "Effective Disaster Management by Efficient Usage of Resources." International Journal of Advanced Engineering, Management and Science 2.12.

[30] Tamayo P,Slonim D, Mesirov J,Zhu Q,Kitareewan S, Dmitrovsky E,Golub T R, (1999), Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, Proceedings of the National Academy of Sciences, vol. 96(6), pp. 2907-2912.

[31] Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang, Lei Hua ,(2004), An Introduction to Cluster Analysis for Data Mining, Journal of Medical Systems, Springer,vol. 36, pp. 2431-2448.

[32] Richard C Dubes, Anil K Jain, (1988), Algorithms for Clustering Data, Prentice Hall,pp.320.

[33] Estivill-Castro V,Yang J, (2000), Fast and robust general purpose clustering algorithms. In PRICAI Topics in Artificial Intelligence , Springer Berlin Heidelberg ,pp. 208-218.

[34] Fraley C, Raftery A E, (1998), How many clusters? Which clustering method? Answers via model-based cluster analysis, The computer journal,vol. 41(8), pp. 578-588.