Middle-East Journal of Scientific Research 24 (7): 2327-2332, 2016
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2016.24.07.23400

Hierarchical Clustering to Determine Missing Values of Yeast Gene

[1]Akey Sungheetha, [2]R. Rajesh Sharma, [1]K. Narmatha, [1]C. Vijesh Joe, [3]S. Iniya Shree and [1]P.D. Mahendhiran

[1]Department of Information Technology, Karpagam College of Engineering,
[2]Department of Information Technology, Hindusthan College of Engineering and Technology,
[3]Department of Computer Science, Bannari Amman Institute of Technology, Sathyamangalam,
Coimbatore, Tamilnadu, India

Abstract: The data to cluster does not pass all the input values on filtering data and hence missing values are identified. The problem of identifying missing values in yeast gene dataset is resolved using hierarchical clustering for the whole yeast gene data. The proposed calculation and analysis on identifying missing values are figured out using the two sets namely small and large yeast gene expression dataset. Both the small and large yeast gene expression dataset samples are processed and compared for clustering methods: single linkage, complete linkage, average linkage and centeroid linkage. The performance metrics correlation and distance are determined for these four methods to prove the best clustering range.

Key words: Cluster · Yeast data · Hierarchical clustering · K means clustering and Filtering data

INTRODUCTION

The behaviour of Loading, Filtering and Adjusting Data in clustering operation is illustrated with conditions. The four clustering methods namely centroid linkage, single linkage, complete linkage and average linkage measures the performance of seven metrics such as Correlation (uncentered), Correlation (centered), Absolute correlation (uncentered), Absolute correlation (centered), Spearman Rank correlation, Kendall's tau, Euclidean distance and City block distance.

Loading, Filtering and Adjusting Data: An example reading taken for loading, filtering and adjusting data for yeast open reading (YORF) is shown in Table 1. In this cluster input table, row represents genes and column represents samples or observations. It indicates a minimal cluster input file resembles as in [1].

The row with YORF=YAL005C at 2hour in Table 1 contains a missing value. In order to identify the missing value, "Present % >= X" is enabled.

Load Data: Similar data file as in Table 1 is the data file of Eisen is loaded. The data file consists of yeast gene expression which is trained with different analytical methods [2, 3, 4, 5, 6, 7 and 11]. The first 31 row set of values are considered for evaluation and the loaded file contributes 31 rows and 79 columns in the dataset. Firstly, the filter data operations to filter genes are scheduled with enable or disable options as in Table 2.

Dataset of 31 Rows and 79 Columns: he limitations in Table 2 are the default values set to cluster genes containing yeast expression [11]. Secondly, while setting the option and values to filter data, missing values are accordingly identified and filtered based on enabling/disabling operations with respect to gene availability.

Filter Data: n order to filter data, the default value is set in order to read the result. Initially the filter operation is not performed and it is not accepted. Hence they are set to NIL as in Equation 1 and Equation 2.

Apply filter = NIL (1)
Accept filter = NIL (2)

Later on the the Options and the values are set in order to filter genes as given in Table 3.

Corresponding Author: Akey Sungheetha, Department of Information Technology,
Karpagam College of Engineering, Coimbatore, Tamilnadu, India.

Table 1: YORF-Yeast open reading frame

S.No	YORF	Omin	30min	1hour	2hour	4hour
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

Table 2: Limitation of filter data operation for filter genes

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

Table 3: Option and default value for filter data operation to filter genes - 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	Maxvalue-Minvalue>=	2.0

Table 4: Option to enable identifying values > 100-80

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

There are six conditions to be followed while filtering data. They are illustrated as follows:

Condition 1: After applying filter operation for the given dataset with specified options in Table 3, then the value of whole 31 rows passes out of 31 rows without any missing values. It is found that there are no missing values. This is done by identifying values as given in Table 4.

Condition 2: If the filter gene %present >=80 then result is same with no missing data. The filter operation is not further necessary.

Condition 3: If the Standard deviation, SD (gene vector) is enabled, 0 values have passed out of 31 rows. Next all the observed values less than 2.0 (SD) are removed.

Condition 4: Filter gene for at least 1 observation with absolute value, abs(Val) greater than 20,

$$abs(Val)>20 \quad (3)$$

Then 3 rows pass out of 31 rows for the condition in Equation 3.

Condition 5: If the filtered gene for minimum value is subtracted from maximum value, then

$$max\ val - min\ val >= 2.0 \quad (4)$$

Equation 4 shows that, with applied filter the result is 3 rows have passed out of 31.

Condition 6: When filtered gene is for maximum value, then

$$maxval>=20 \quad (5)$$

Equation 5 on applying filter has 21 rows passed out of 31 rows. The filter is accepted for condition 3, 4, 5 and 6 in order to accept filtered rows.

Adjust Data-Units Mean: The data is adjusted interms of log transform data, center gene-mean, center arrays-mean, normalizing gene and normalizing arrays since the center gene and center array order of operation has its median for valuation to adjust data.

Proposed Study on Clustering for Small Sample Set

Hierarchical (Gene) Clustering: The hierarchically connected genes and arrays are clustered and its corresponding weights are calculated [8, 9]. Then the given weight options of cutoff 0.1 and exponent 1 are default values set to calculate weights. While setting the default values, the similarity metric measures shows the correlation is uncentered for calculated weights. Next the centroid linkage method is used to cluster and generate the clustered resultant. The existence of the metric correlation (uncentered) clustering method is discovered, rely on centroid linkage clustering. First gene tree file (.gtr) is generated gene tree with node and gene values with its exponent, then second Atari array tree (.atr) disk image (a copy of 8 bit formatted disk) file is generated with node and its array values with same exponent 1 and third corel draw text editor image template (.cdt) is generated with the E weight (exponent weight) of G weight (Gene Weight). The similar performance for the centroid linkage method in hierarchical clustering is followed for single linkage method, complete linkage method and average linkage method. The centroid linkage method involves 2 node and gene values for generated gene tree in sample 1.

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

Sample 1

The interference for the single linkage method shows generated values in sample 2.

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

Sample 2

The rest of the files are same for centroid linkage and single linkage method. The complete linkage method for clustering differs in value links from others. It generates sample 3 gene tree file.

Node 1x	Gene 1x	Gene 0x	-0.527353
Node 2x	Gene 2x	Gene 1x	-0.819571

Sample 3

For average linkage, gene tree file in sample 4 is generated showing one different value for node 2 similar to other methods of clustering.

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

Sample 4

After performing cluster operation for the generated values using average linkage method, k-means clustering is chosen for evaluation. The similar dataset of Eisen which is fed for hierarachical clustering is used in k-means clustering.

K-Means (Gene) Clustering Technique: The genes and arrays for k-mean clustering organize genes and arrays respectively. Both have 10 numbers of cluster k and 100 numbers of runs each. The two methods that can be performed are: k-means and k-medians. On execution of k-means with given similarity metric, Euclidean distance for both gene and array; it is found that clusters are available more in number than the genes. Entire dataset is passed without any gene filter irrespective of number of observations or absolute value specification. The data is adjusted and it is independent of hierarchical technique. After execution the cluster k generate cluster gene file (.kgg) where gene groups 10 clusters k [10, 11 and 12].

The data in open reading frame (ORF) is a .kgg file and .kag file. It groups the gene into 10 groups. Cluster, k for 10 gene and 10 arrays lists the gene weight and experiment weight.

Self-Organized Mapping and Principle Component Analysis: After execution of k-means clustering technique, the same Eisen dataset is tested in Self organized mapping (S0M). The SOM made reaches in calculating it with genes and Arrays by initially organizing genes and arrays respectively similar to k-means clustering. The X dimension and Y dimension for the genes and arrays are set 3. The number of iterations for genes by default is 1, 00, 000 and arrays have 20 000 respectively. The initial tau is set to 0.02 by default for calculation. Both genes and arrays of SOM have the same value. The similarity metric here is the Euclidean distance common for both. Three files generated of which GNF file show the gene vectors and ANF file show the array vectors. The gene/array file together shows the gene weight and experiment weight of the vectors. The means values are not presented in self o rganized maps [13]. 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 coordinate 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. Data must be filtered before adjusting process.

Filter Data: Filtering data with mean is similar to filtering 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 0 passed out of 31. Adjusting median first and then filtering also shows 0 passed out of 31. Initial step to filter gene atleast 1 observation with abs(val)>=2.0 on applying filter gives 3 passed out of 31. The filter is accepted to perform clustering. Adjusting the data for the center gene and center array to mean and median respectively and vice versa filter 0 passed out of 31.

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. The difference lines in selecting the median instead of mean on center gene or center arrays respectively.

Table 5: The tabulated results for all 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

Proposed Study on Clustering for Hierarchical (Gene) Clustering Technique - Large Sample Set: The similarity metric type's performances are measured. The various similarity metrics used in gene clustering are: Correlation (uncentered), Correlation (centered), Absolute correlation (uncentered), Absolute correlation (centered), Spearman Rank correlation, Kendall's tau, Euclidean distance and City block distance.

Table 5 gives a comparison idea on the measure of gene and array values during clustering. Also it helps in identifying the missing values of yeast gene.

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).

The codes for the methods are represented as follows:

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

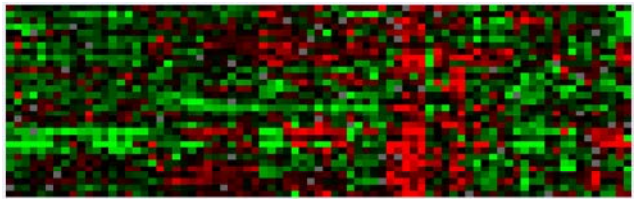

Fig. 1: Gene tree view 31rows 79columns

Table 6. 31rows and 2467 rows (79 columns) – cluster range

Cluster method	Range	31 rows node/gene	31rows 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	-0.28906	0.0223

Table 8: 31rows and 2467 rows (79 columns) – execution time

	Time (sec)			
Cluster method	31 rows node/gene	31rows 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

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

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 method are tested for all the other similarity metrics and the performance is updated in Table 1. For correlation centered, the corresponding procedure code H_GA_C_CC_CEL, H_GA_C_CC_SL, H_GA_C_CC_COL and H_GA_C_CC_AL. 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).

The previous discussions involve the observations for 31rows 79columns. On increasing the size to 2467 rows 79 columns as given in Table 6, clustering performance is maintained with effective clustering ways such as eucledian and city block distance approach and hence the performance comparision is done with large dataset [14, 15, 16 and 17]. The results obtained are tabulated for node/gene and node/array respectively. The time taken for the process of performing ACC, SRC and KT clustering increases gradually using gene cluster tool. The ACU, ACC, ED and CBD gene/array values involve cluster existing positive values. When weight has cutoff=0.1 and exponent=1 for gene and arrays, 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 (CEL) ie some value resemble same.

In case of execution time ED and CBD method takes very less duration to process the data as given in Table 8.

CONCLUSION

Similar to CU, the SR for CC, ACU, ACC, SRC and KT 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. Xutao Deng, Omaha, Geng, H. Ali, H., 2005. "Learning yeast gene functions from heterogeneous sources of data using hybrid weighted Bayesian networks", Computational Systems Bioinformatics Conference, 2005. IEEE Proceedings, pp: 25-34. 8-11 Aug. 2005.
2. Sanghamitra Bandyopadhyay, Anirban Mukhopadhyay, Ujjwal Maulik, "An Improved Algorithm for Clustering Gene Expression Data, 23(21): 2859-2865.
3. Rao, A., 2002. "A clustering algorithm for gene expression data using wavelet packet decomposition", Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers, 1: 316-319.
4. Tseng, G.C., 2004. "A comparative review of gene clustering in expression profile", ICARCV 2004 8th Control, Automation, Robotics and Vision Conference, 2: 1320-1324.
5. Chi Kin Chow, 2009. "A cooperative feature gene extraction algorithm that combines classification and clustering", IEEE International Conference on Bioinformatics and Biomedicine Workshop, pp: 197-202.
6. Dutta, D., 2012. "A genetic weighted k-means algorithm for clustering gene expression data", 2012 1st International Conference on Recent Advances in Information Technology (RAIT), pp: 548-553, 15-17 March 2012.
7. Choudhury, N., 2012. "A modified QT-clustering algorithm over Gene Expression data", 2012 1st International Conference on Recent Advances in Information Technology (RAIT), pp: 542-547, 15-17 March 2012.
8. Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58: 236-244.
9. Murtagh, F., 1984. A survey of recent advances in hierarchical clustering algorithms which use cluster centers. Comput. J., 26: 354-359.
10. Leonard Kaufman and Peter J. Rousseeuw, 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons.
11. Eisen, M.B., PT. Spellman, P.O. Brown and D. Botstein, 1998. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci. USA, 95: 14863-68.
12. Tavazoie, S., J.D. Hughes, M.J. Campbell, R.J. Cho and G.M. Church, 1999. Systematic determination of genetic network architecture. Nat Genet., 22: 281-85.
13. Tamayo, P., D. Slonim, J. Mesirov, *et al*.., 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci. USA, 96: 2907-12.
14. Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang and Lei Hua, 2004. " An Introduction to Cluster Analysis for Data Mining".
15. Richard C. Dubes and Anil K. Jain, 1988. Algorithms for Clustering Data, Prentice Hall.
16 Estivill-Castro, V. and J. Yang, 2000. A Fast and robust general purpose clustering algorithm. Pacific Rim International Conference on Artificial Intelligence, pp: 208-218.
17. Fraley, C. and A.E. Raftery, 1998. "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis", Technical Report No. 329. Department of Statistics University of Washington, 1998.