International Journal of Biomedical Materials Research
2017; 5(5): 59-63
http://www.sciencepublishinggroup.com/j/ijbmr
doi: 10.11648/j.ijbmr.20170505.11
ISSN: 2330-7560 (Print); ISSN: 2330-7579 (Online)

On the Selection of Appropriate Proximity Measurement for Gene Expression Data

Md. Bipul Hossen[1, *], Arefin Mowla[1], Md. Harun or Rashid[1], Md. Binyamin[2]

[1]Department of Statistics, Begum Rokeya University, Rangpur, Bangladesh

[2]Department of Statistics, Mawlana Bhashani Science and Technology University, Santosh, Tangail, Bangladesh

Email address:

mbipu.ru@gmail.com (Md. B. Hossen), arefinmowla.milu@gmail.com (A. Mowla), hb0910009@gmail.com (Md. H. or Rashid), rony4721@gmail.com (Md. Binyamin)

[*]Corresponding author

To cite this article:

Md. Bipul Hossen, Arefin Mowla, Md. Harun or Rashid, Md. Binyamin. On the Selection of Appropriate Proximity Measurement for Gene Expression Data. *International Journal of Biomedical Materials Research.* Vol. 5, No. 5, 2017, pp. 59-63. doi: 10.11648/j.ijbmr.20170505.11

Received: January 28, 2017; **Accepted**: February 17, 2017; **Published**: June 30, 2017

Abstract: Gene expression profile has become a useful biological resource in recent years and its plays an important role in a broad range of biology. But a large number of genes and the complexity of biological networks greatly increase the evaluation of comprehending and interpreting the resulting mass of data, which often consists of millions of measurements. In the computational analysis of gene expression data, the main aspect is to finding co-expressed genes as the proximity (similarity or dissimilarity) measures that are used in the clustering method. Several number of proximity measures work are used in the gene data but the majority of these works has given emphasis on the biological results and no critical assessment of the suitability of the proximity measures for the analysis of gene expression data. For these consequences this paper is to investigate the appropriate proximity measurement for gene expression data. As a case study, we considered six real datasets. Based on this, we provide a comparative study of five proximity measures: Euclidean distance, Manhattan distance, Pearson correlation, Spearman correlation, Cosine distance. We discuss Adjusted Rand Index, Silhouette Index of clustering to assess the quality and reliability of the results. Our results reveal that the Cosine distance method with complete linkage exhibited the best performance for both Affymetrix and cDNA datasets according to Adjusted Rand Index. Our results also reveal that the Spearman correlation measure with complete linkage exhibited the best performance for both Affymetrix and cDNA datasets according to Silhouette Index.

Keywords: Proximity Measures, Agglomerative Hierarchical Clustering, Adjusted Rand Index, Silhouette Index, Gene Expressions Data

1. Introduction

Microarray technology measures the evolution of thousands of genes quantitatively and simultaneously in a gene expression profiling experiment under different [1]. An appropriate proximity measure is highly demanded to extract hidden information from co-expression analysis of enormous genome data. In that case, a common task is to compare the proximity measures for gene expression datasets. DNA microarray technology has now made it possible to simultaneously monitor the evolution levels of thousands of genes during important biological processes and across collections of related samples.

There are several widely used proximity measures, such as Euclidean Distance, Manhattan Distance, Cosine Distance, Pearson Correlation, Spearman Correlation, Jaccard Coefficient, Kendall Tau Correlation Coefficient etc. Besides, various analytical and statistical approaches are already developed to capture the overall feature of high dimensional variable datasets. Hierarchical clustering method is one of them, which is classified into agglomerative hierarchical methods and divisive hierarchical methods. Agglomerative Hierarchical Clustering (AHC) is more popular between them. There are several AHC methods are well established [2, 3].

Single channel microarrays (Affymetrix) and double channel microarrays (cDNA) are two types of platforms where the gene expression microarray technology is available and these datasets are meaningful to cluster both genes and samples [4, 5, 6]. The above types datasets are usually used for

gene based clustering and sample based clustering. But this study conducted only sample based clustering because the goal of sample-based clustering is to identify the phenotype structures or substructures of the samples. In the sample based clustering, genes are treated as features while samples are treated as objects and samples are partitioned into homogeneous groups.

In this study, five proximity measures (Euclidean Distance, Manhattan Distance, Cosine Distance, Pearson Correlation, and Spearman Correlation) are used to identify the clustering performance in gene expression [7, 8]. Four AHC methods (Single Linkage, Complete Linkage, Average Linkage and Centroid Linkage) were discussed in [8, 9, 10] which are used to identify the clustering performance in gene expression data.

Four AHC methods (Single Linkage, Complete Linkage, Average Linkage and Centroid Linkage) were accomplished to evaluate the clustering performance in their analysis are expressed [11, 12, 13]. However most of the author's demonstrated cosine correlation method is better and rest of the author's demonstrated Euclidean distance is better measure to evaluate microarray gene expression data in their analysis.

1.1. Proximity Measures of Gene Expression Data

Proximity measures (distances and similarities) are supplementary material for gene expression data analysis are analysis by these two author [14, 15]. For this reason we introduce some proximity measures (distance and similarity) here. Suppose x and y be denoted as two numerical vector of gene expression data objects with m features, where the object can be either genes or samples are detailed in [16, 17, 18]. Then the measures (Euclidean Distance Method, Manhattan Distance Method, Cosine Distance Method, Pearson Correlation Measure, and Spearman Correlation Measure) can be expressed in [19, 20, 21, 22] that are given below.

1.1.1. Euclidean Distance

The distance between x and y is the square root of squared difference between corresponding elements of the two vectors. It can be defined as

$$d(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$$

1.1.2. Manhattan Distance

The distance between x and y is measured along axes at right angles and it is defined as

$$d(x,y) = \sum_{i=1}^{m}|x_i - y_i|$$

1.1.3. Cosine Similarity

Cosine similarity is widely used similarity measure applied to text documents, such as in numerous information retrieval applications and clustering too. Cosine similarity is popular because it is efficient to evaluate, especially for sparse vectors, as only the non-zero dimensions need to be considered. The independency of document length is an important property of cosine similarity. Therefore, the cosine similarity ignores 0-0 matches like the Jacquard measure. The cosine similarity is defined by the following equation.

$$d(x,y) = \frac{x \times y}{\|x\| \times \|y\|}$$

1.1.4. Spearman Correlation

Spearman measures the degree of a monotonic relationship between two variables, without making any assumptions about the frequency distribution of the variables. In practice, a simple formula is normally used to calculate Spearman Correlation.

$$cor(x,y) = 1 - \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{m}(x_i - \bar{x})\hat{}2 \sum_{1}^{m}(y_i - \bar{y})\hat{}2}$$

1.1.5. Pearson Correlation

Pearson correlation coefficient is widely used and has proven effective as a similarity measure for gene expression data. Pearson correlation is defined by the following equation.

$$cor(x,y) = \frac{Cov(x,y)}{SD(x) \times SD(y)}$$

Where, COV is the covariance between x and y, SD is the standard deviation.

1.2. Checking Validity

On the selection of appropriate proximity measures it is common to evaluate the result of those measures with clustering. But clustering is an unsupervised process in the data mining and pattern recognition and most of the clustering methods are very impressible to their input parameters. Therefore it is very important to evaluate the result of the clustering methods. It is difficult to characterize when a clustering result is acceptable, thus several clustering validity techniques have been well developed. In this study the most commonly used validity techniques- Adjusted Rand Index and Silhouette Index are used.

1.2.1. Adjusted Rand Index (ARI)

For cases in which a mention partition is available one can employ emerged validation measures to foretell the quality of the results. Due to its emendation that takes into account equivalencies between partitions [23]. We choose the Adjusted Rand is discussed at Bipul [19], which is defined as given below for the evaluation of clustering results. The greater its value, the greater is the resemblance between the two partitions under comparison, with values close to 0 representing equivalencies found by chance. Given a partition U and a mention partition V, (a) accounts for the total number of item pairs belonging to the same cluster in both U and V; (b) represents the total number of object pairs in the same cluster in U and in different clusters in V; (c) is the total number of object pairs that are in different clusters in U and in the similar cluster in V; and (d) is the total number of object pairs that are

in dissimilar clusters in both U and V.

$$ARI = \frac{a - \frac{(a+b)(a+c)}{(a+b+c+d)}}{\frac{(a+b)(a+c)}{2} - \frac{(a+b)(a+c)}{(a+b+c+d)}}$$

1.2.2. Silhouette Index (SI)

To invoice the number of clusters in our third amends view, a corresponding index of balance between partitions is also devoted. The Silhouette index is defined as considering a partitioning of m objects in k disjoint clusters. Here, the average distance among x and all the left over objects of its cluster is represented by u(i). On the other hand, for a conferred object x, the usual distance of x and all the other objects from a given cluster is obtained and is denoted by v(i). This process is repeated for all the k–1 clusters, excluding the cluster belongs to x. At the end of the scheme the lowest average value found is assigned to v(i). In a single words, the mean distance between x and its adjacent cluster (closest cluster) is denoted by v(i). Silhouette, which is a maximization measure, has its values within [−1, 1].

$$s = \frac{1}{m}\sum_{i=1}^{m}\frac{v(i) - \mu(i)}{\max\{v(i), \mu(i)\}}$$

We choose the Silhouette based on its superior consequences in comparison to other relative criteria [24]. We also message that the Silhouette has already been successfully employed in order to estimate the number of cluster for gene expression data.

Finally, it can be noted that by using the SI one can simulate a real application in which the user need not any a priori information regarding the number of clusters present in the data. It is significant to make clear, that the use of comparative indexes (such as the Silhouette) is just part of the more general procedure that comprehends the entire clustering analysis.

Tendentious by this problem it is momentous to envisage all of the methods for gene data by standardized which method are relatively best. In this paper, it is tried to compare five proximity measures for the both Affymetrix and cDNA datasets. It is also provided a detailed graphical as well as analytical comparison. We used Bar diagram as well as ARI and SI to check the suitable proximity measures for clustering. This paper is prepared by using the AHC algorithm with several proximity measures are redacted using language programming R 3.0.2. Several times Ms-Excel and Ms-Word are used as calculation and typing software.

2. Experiments and Results

There are six publicly available microarray datasets from [9] which are related to our analysis. These datasets can be classified into single channel as Affymetrix chip (3 sets) and double-channel as cDNA (3 sets). We compare five proximity measures with four different clustering methods. Generally the gene expression data set is so much noisy, concurrence with expression pattern, beneath constitutional and up constitutional so it is essential to take preprocess before differential analysis. To adjust data for technical segment, as averse to biological differences between the samples we have preprocessed only Affymetrix data by using standardized procedure. It is noted that the cDNA datasets were preprocessed. The empirical datasets are given in Table 1, where n is the number of sample, m is the number of feature as genes and d is the number of feature after filtering.

Table 1. *Description of Affymetrix and cDNA datasets.*

Dataset	Chip	Tissue	n	#c	Dist. Classes	m	d
Armstrong-V2	Affy	Blood	72	3	24,20,28	12582	2194
Bhattacharjee	Affy	Lung	203	5	139,17,6,21,20	12600	1543
Nutt-V1	Affy	Brain	50	4	14,7,14,15	12625	1377
Alizadeh-V2	cDNA	Blood	62	3	42,9,11	4022	2093
Bredel	cDNA	Brain	50	3	31,14,5	41472	1739
Garber	cDNA	Lung	66	4	17,40,4,5	24192	4553

Firstly we present some graphical displays for both gene expression datasets. For each of the five proximity measures along with four AHC methods of clustering, we embody the results by using Bar diagram, to compare which proximity measures is meaningful and the results are given in Figure 1, Figure 2 Figure 3 and Figure 4.

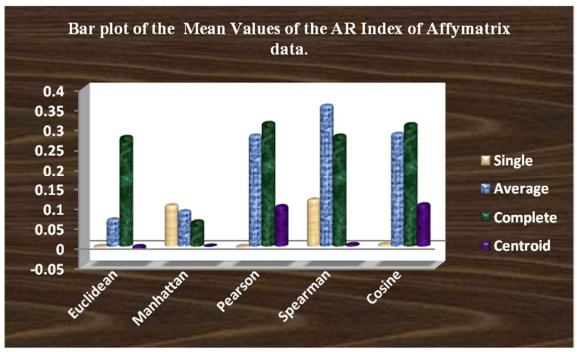

Figure 1. *Bar plot of the Mean of the AR Index of Affymetrix data.*

The mean values of the Adjusted Rand index of the experiments with Affymetrix datasets are presented in Figure 1. The cosine method with respect to complete linkage obtained the maximum value with respect to AHC methods when compared to those achieved by the other methods.

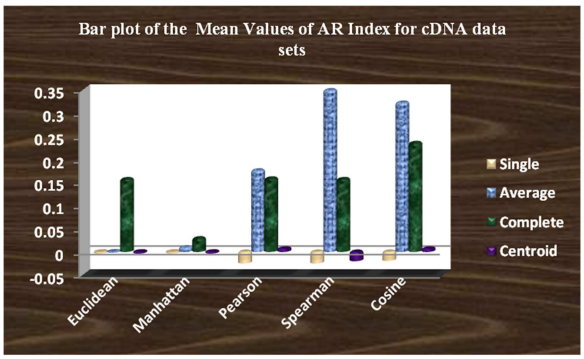

Figure 2. *Bar plot of the Mean of the AR Index of cDNA data.*

Figure 2 illustrates the mean values of the Adjusted Rand for the experiments performed with the cDNA datasets. The

cosine method achieved the highest value with respect to proximity measures in comparison to all the other methods. Therefore the cosine methods and complete linkage give the best result in comparison to all the other methods.

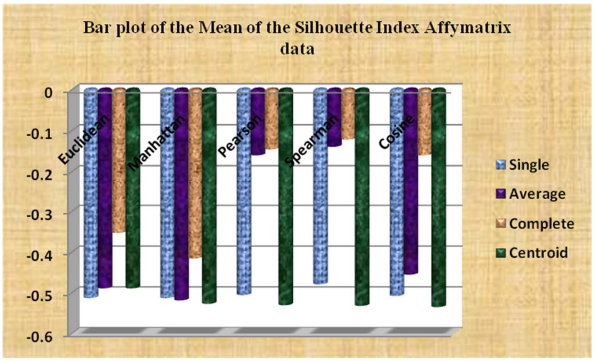

Figure 3. *Bar plot of the Mean Values of the Silhouette Index of Affymetrix data.*

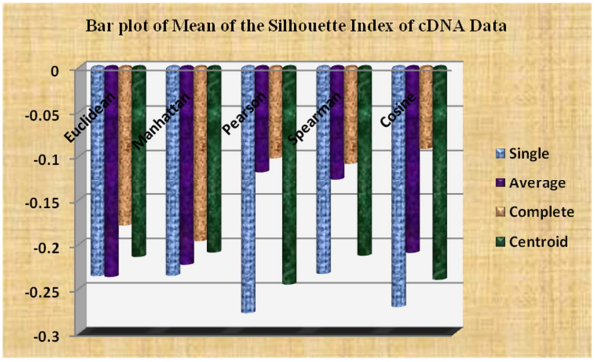

Figure 4. *Bar plot of the Mean Values of the Silhouette Index of cDNA data.*

The mean values of the Silhouette Index of the experiments with Affymetrix datasets are presented in Figure 3. The Spearman method obtained the maximum value with respect to AHC methods when compared to those achieved by the other methods.

Figure 4 illustrates the mean values of the Silhouette Index for the experiments performed with the cDNA datasets. The method achieved the highest value with respect to proximity measures in comparison to all the other methods. The Spearman methods and complete linkage give the best result in comparison to all the other methods.

Table 2. *The mean adjusted Rand value of Affymetrix and cDNA datasets.*

Affymetrix datasets				
	Single Linkage	**Average Linkage**	**Complete Linkage**	**Centroid Linkage**
Euclidean	-0.0037	0.0672	0.2771	-0.0077
Manhattan	0.1025	0.0874	0.0622	-0.0009
Pearson	-0.0034	0.2803	0.3121	0.1005
Spearman	0.1172	0.3558	0.2805	0.0049
Cosine	0.005	0.2849	0.3094	0.1049
cDNA datasets				
Euclidean	-0.0035	-0.0019	0.1580	-0.0035
Manhattan	-0.0035	0.0092	0.0287	-0.0035
Pearson	-0.0247	0.1769	0.1595	0.0065
Spearman	-0.0247	0.3494	0.1582	-0.0199
Cosine	-0.0193	0.3225	0.2359	0.0065

Table 3. *The mean silhouette index value of Affymetrix and cDNA datasets.*

Affymetrix datasets				
	Single Linkage	**Average Linkage**	**Complete Linkage**	**Centroid Linkage**
Euclidean	-0.5124	-0.4886	-0.3524	-0.4886
Manhattan	-0.5126	-0.5181	-0.4150	-0.5266
Pearson	-0.5051	-0.1598	-0.1454	-0.5300
Spearman	-0.4789	-0.1391	-0.1210	-0.5316
Cosine	-0.5074	-0.4549	-0.1599	-0.5346
cDNA datasets				
Euclidean	-0.2357	-0.2366	-0.1786	-0.2140
Manhattan	-0.2352	-0.2228	-0.1962	-0.2191
Pearson	-0.2774	-0.1178	-0.1020	-0.245
Spearman	-0.2331	-0.1259	-0.1083	-0.2126
Cosine	-0.2704	-0.2094	-0.0914	-0.2397

Table 2 shows the the mean values of Adjusted Rand Index to check the performance of the proximity measures along with the AHC methods. For Affymetrix datasets cosine with complete linkage gives the best result and for the cDNA datasets cosine gives the partially best result. The overall analysis gives the Cosine distance method with complete linkage exhibited the highest result according to Adjusted Rand Index.

The mean values of Silhouette Index of 5 proximity measures with 4 clustering methods for both Affymetrix and cDNA datasets are presented in Table 3 to observe the best proximity measure with respect to clustering methods. The spearman correlation method with complete linkage shows on average highest values according to Silhouette Index for both types of datasets.

3. Conclusion

We show here a comparative study of five proximity measures with four clustering algorithms applied on six clinical cancer gene expression datasets. Our results reveal that the Cosine distance method with complete linkage exhibited the best performance for both Affymetrix and cDNA datasets according to Adjusted Rand Index. This analysis also shows the Spearman Correlation method with complete linkage exhibited the best performance for both Affymetrix and cDNA datasets according to Silhouette Index. Additionally, among the clustering methods the complete linkage gives the best result according to ARI and SI for both types of datasets. To the best of our knowledge, the comparative study of proximity measure with the validity index as Adjusted Random Index and Silhouette Index are poorly documented in literature.

References

[1] Brown M P and Bostein D (1999); Exploring the new world of genome with DNA microarrays. Nature Genetics, vol. 21 (1), pp. 33-37.

[2] Cunningham K M and Ogilvie J C (1972); Evaluation of hierarchical grouping techniques: A preliminary study. The Computer Journal, vol. 15 (3), pp. 209–213.

[3] Johnson R A and Wichern D W (2002). Applied Multivariate Statistical Analysis. Upper Saddle River, NJ: Prentice Hall.

[4] Monti S, Tamayo P, Mesirov J, Golub T (2003); Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data; Machine Learning. Vol. 52 (1), pp. 91-118.

[5] Daxin J, Chun T, and Aidong Z (2004); Cluster Analysis for Gene Expression Data: A Survey, IEEE Transactions on Knowledge and Data Engineering, vol. 16 (11), pp. 1370-1386.

[6] Costa I G, Carvalho F A D and Souto M C P D (2004); Comparative Analysis of Clustering Methods for Gene Expression Time Course Data. Genetics and Molecular Biology, vol. 27 (4), pp. 623-631.

[7] Kerr G, Ruskin H J, Crane M and Doolan P (2008); Techniques for clustering gene expression data. ComputBiol Med, vol. 38 (3), pp. 283-293.

[8] Geetha T and Michael A (2010); Enhanced Hierarchical Clustering for Gene Expression data. International Journal of Computer Applications, vol. 1 (20), pp. 92–98.

[9] Marcilio C P de Souto, Ivan G Costa, Daniel S A de Araujo, Teresa B Ludermir and Alexander Schliep (2008); Clustering cancer gene expression data: a comparative study. BMC Bioinformatics, pp. 01-14.

[10] Kuiper F K and Fisher L (1975); A Monte Carlo comparison of six clustering procedures. Biometrics, vol. 31 (8), pp. 777–783.

[11] Hubert L (1974); Approximate evaluation techniques for the single-link and complete link hierarchical clustering procedures. Journal of the American Statistical Association, vol. 69, pp. 698–704.

[12] Blashfield R K (1976); Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. The Psychological Bulletin, vol. 83, pp. 377–388.

[13] Hands S and Everitt B (1987); A Monte Carlo study of the recovery of cluster structure in binary data by hierarchical clustering techniques. Multivariate Behavioral Research, vol. 22 (2), pp. 235–243.

[14] Anderberg M (1973); Cluster analysis for applications. New York: Academic Press.

[15] Jain A K and Dubes R C (1988); Algorithms for clustering data, Prentice Hall.

[16] Guojun G, Chaoqun M and Jianhong W (2007); ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA. Data Clustering: Theory, Algorithms, and Applications

[17] Gentleman R, Ding B, Dudoit S and Ibrahim J (2005); Bioinformatics and Computational Biology Solutions Using R and Bioconductor Statistics for Biology and Health, Springer-Verlag London Limited.

[18] Pablo A Jaskowiak, Ricardo J G B Campello and Ivan G Costa (2013); Proximity Measures for Clustering Gene Expression Microarray Data: A Validation Methodology and a Comparative Analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 10 (4), pp. 845-857.

[19] Md. Bipul Hossen, Md. Siraj-Ud-Doulah, Aminul Hoque (2015); Methods for Evaluating Agglomerative Hierarchical Clustering for Gene Expression Data: A Comparative Study, Computaitonal Biology and Bioinformatics, Vol. 3 (6), pp. 88-94.

[20] Md. Siraj-Ud-Doulah, Md. Bipul Hossen (2016); Performance Evaluation of Clustering Methods in Microarray Data. American Journal of Bioinformatics Research, Vol. 6 (1), pp. 19-25.

[21] Jaskowiak P A, Campello R J G B and Costa I G (2013); Proximity Measures for Clustering Gene Expression Microarray Data: A Validation Methodology and a Comparative Analysis, Computational Biology and Bioinformatics. Vol. 10 (4), pp. 845-857.

[22] Eldesoky, A. E, M. Saleh, N. A. Sakr (2009); Novel Similarity Measure fo Document Clustering Basedon Topic Phrase, International Conferenceon Networking and Media Convergence, vol. 24, pp. 92-96.

[23] Milligan G W and Cooper M C (1988); A study of standardization of variables in cluster analysis. Journal of Classification, vol. 5 (2), pp. 181-204.

[24] Peter J. Rousseeuw (1987); Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics. Vol. 20: pp. 53–65.