Performance Evaluation of Distance Metrics in the Clustering Algorithms

VIJAY KUMAR[1]
JITENDER KUMAR CHHABRA[2]
DINESH KUMAR[3]

[1]Computer Science & Engineering Department, Manipal University, Jaipur, Rajesthan, INDIA
[2]Computer Engineering Department, National Institute of Technology, Kurukshetra, Haryana, INDIA
[3]CSE Department, Guru Jambheshwer University of Science & Technology, Hisar, Haryana, INDIA
[1]vijaykumarchahar@gmail.com
[2]jitenderchhabra@gmail.com
[3]dinesh_chutani@yahoo.com

Abstract. Distance measures play an important role in cluster analysis. There is no single distance measure that best fits for all types of the clustering problems. So, it is important to find set of distance measures for different clustering techniques on datasets that yields optimal results. In this paper, an attempt has been made to evaluate ten different distance measures on eight clustering techniques. The quality of the distance measures has been computed on basis of three factors: accuracy, inter-cluster and intra-cluster distances. The performance of clustering algorithms on different distance measures has been evaluated on three artificial and six real life datasets. The experimental results reveal that the performance and quality of different distance measures vary with the nature of data as well as clustering techniques. Hence choice of distance measure must be done on basis of dataset and clustering technique.

Keywords: Distance Measures; Clustering Algorithms; Ant Colony based Clustering; Modified Harmony Search Clustering.

(Received May 26th, 2014 / Accepted August 8th, 2014)

1 Introduction

Clustering is an important data mining technique where information about labeling and structure is not available. It is the process of partitioning a set of data points into different groups such that the data in each group are similar to each other. Clustering algorithms are broadly classified into two groups: hierarchical and partitional [7]. Hierarchical clustering algorithms recursively find nested clusters either in agglomerative mode or in divisive mode. The former one starts with each data point in its own cluster and merges the most similar pair of clusters successively to form a cluster hierarchy and the latter starts with all the data points in one cluster and recursively divides each cluster into smaller clusters [21]. The well-known agglomerative hierarchical clustering algorithms are single, average, complete and weighted linkage. On the other hand, partitional clustering groups the data points into some pre-specified number of clusters without using hierarchical structure. The most popular partitional clustering techniques are K-Means, K-Medoid, Fuzzy C-Means and Expectation-Maximization.

These clustering techniques are based on similarity between data points, which is determined by a distance measure. The distance measure plays an important role in obtaining correct clusters. The selection of right distance measure affects the results of clustering algorithms. They may affect the shape, volume and orientation of clusters as some data points may be close to one another according to one distance measure and far way according to another distance [19, 20]. The motivation of this paper is to analyze the effect and evaluate the performance of various distance measures on different clustering techniques.

In this paper, we study the distance measures from a new perspective: how they affect the clustering results. The ten well-known distance measures are discussed with their relative strengths and weaknesses. These are: Euclidean, Standardized Euclidean, Manhattan, Mahalanbois, Cosine Similarity, Pearson Correlation, Spearman Correlation, Chebychev, Canberra, and Bray-Curtis. These are evaluated in conjunction with eight different clustering techniques over nine different datasets. The rest of the paper is organized as follows. Section 2 presents clustering techniques. Section 3 introduces distance measures that used for numerical data sets in clustering. In Section 4, the effect of distance measures on clustering techniques is investigated. Finally, a concluding remark is drawn in Section 5.

2 Clustering Techniques

The clustering algorithms are used to partition the dataset $X = x_1, x_2, \ldots, x_j, \ldots, x_N$, where $x_j = (x_{j1}, x_{j2}, \ldots, x_{jd}) \in R^d$ into a number of clusters, say K, $(C_1, C_2, \ldots, C_K)$. The parition matrix $U(X)$ is represented as $U = [u_{kj}], k = 1, 2, \ldots, K$, and $j = 1, 2 \ldots, N$, where u_{kj} is membership of datapoint x_j to clusters C_k. The $u_{kj} = 1$ if $x_j \in C_k; otherwise, u_{kj} = 0$.

2.1 Hierarchical Clustering Techniques

The agglomerative hierarchical clustering techniques have been used in this paper. The well- known agglomerative hierarchical techniques are single linkage, average linkage, complete linkage and weighted linkage.

The single linkage clustering is based on the local connectivity criterion [7]. It is also known as a nearest neighbor method. It starts by considering each data point in a cluster of its own. It computes the distances between two clusters p and q such as [13]

$$D_{SL}(p, q) = \min_{x_i \in p, x_j \in q} \{d(x_i, x_j)\} \qquad (1)$$

Based on these distances, it merges the two closest clusters and replacing them by one merged cluster. The distances of the remaining clusters from the merged cluster are recomputed as mentioned above. This process continues until all the data points are in a single cluster. The main advantage of single linkage is that it can handle non-elliptical shapes. However, it is sensitive towards noise and outliers [7, 18].

The average linkage clustering has a similar procedure as the single linkage except the distance computation between two clusters. It uses the average of pairwise distance between points in two clusters p and q as:

$$D_{AL}(p, q) = \frac{1}{|p||q|} \sum_{x_i \in p} \sum_{x_j \in q} d(x_i, x_j) \qquad (2)$$

It is less susceptible to noise and outliers. The one disadvantage is its biasing towards globular clusters [18].

The complete linkage clustering is also called the furthest method. It uses the largest distance between data points in two clusters p and q as:

$$D_{CL}(p, q) = \max_{x_i \in p, x_j \in q} \{d(x_i, x_j)\} \qquad (3)$$

It does not account for cluster structure. It cannot detect the non-spherical clusters. The weighted average linkage method is also known as weighted pair group method using arithmetic average. The difference between average and weighted linkage is that the distances between the newly formed cluster and the rest are weighted based on the number of data points in each cluster.

2.2 Partitional Clustering Techniques

The well-known partitional techniques are K-Means and K-Mediods. The main disadvantages of these techniques are that these are easily trapped in local optima. The K-Means is well-known partitional clustering algorithm [7]. It seeks an optimal partition of data by minimizing the sum-of-squared-error criterion with an iterative optimization procedure such as [7, 13]

$$J(U, V) = \sum_{j=1}^{N} \sum_{i=1}^{K} u_{ij} \|x_j - v_i\|^2 \qquad (4)$$

where v_i is the center of cluster C_i. Here, cluster centers are initialized by randomly chosen data points form dataset. Each data point is assigned to the nearest cluster using minimum distance criterion. Thereafter, the cluster centers are updated to the mean of data points belonging to them. This process is repeated until there is no change for each cluster. The disadvantage of K-Means is that it is sensitive towards initialization of cluster centers.

The K-Medoid algorithm is an adaptation of K-Means algorithm. Rather than calculating the mean of data points in each cluster, medoid is chosen for each cluster at each iteration.

Shelokar et al. [17] described an ant colony optimization methodology for data clustering (ACOC). It mainly

relies on pheromone trails to guide ants to group data points according to their similarity and on a local search that randomly tries to improves the best iteration solution before updating pheromone trails.
Kumar et al. [9, 10] developed a modified harmony search based clustering ($MHSC$) technique. Here cluster center based encoding scheme is used. Each harmony vector contains K cluster centers, which are initialized to K randomly chosen data points from the given dataset. This process is repeated for each of the HMS vectors in the harmony memory, where HMS is the harmony memory size. The data points are assigned to different cluster centers based on minimum Euclidean distance criterion and cluster centers represented by the harmony vectors are replaced by the mean data points of respective clusters. The fitness of each harmony vectors is computed using sum-of-squared-error criterion and is minimized using modified harmony search algorithm. The improvisation process is used to update the harmony vectors. In $MHSC$, the processes of fitness computation and improvisation are executed for a maximum number of iterations. The best harmony vector at the end of last iteration provides the solution to the clustering problem.

3 Distance Measures

The distance measure must be determined before the clustering. It reflects the degree of separation among target data points and should correspond to the characteristics used to distinguish the clusters embedded in the dataset [6, 3]. These characteristics are data dependent in most of cases. There is no single distance measure that is best for all types of clustering problems. Therefore, understanding the importance of different distance measures will help us to choose the best one. Every distance measure is not a metric. To qualify as a metric, a measure must satisfy the following four conditions [7, 21].

1. The distance between any two data points must be non-negative, i.e.,
   $D(x_i, x_j) \geq 0$ for all x_i and x_j
2. The distance between two data points must be zero if and only if the two data points are identical, i.e.,
   $D(x_i, x_j) = 0$ if and only if $x_i = x_j$
3. The distance from x_i to x_j is the same as the distance from x_j to x_i, i.e.,
   $D(x_i, x_j) = D(x_j, x_i)$
4. The distance measure must satisfy the triangle inequality, which is
   $D(x_i, x_j) + D(x_j, x_k) \geq D(x_i, x_k)$ for all x_i, x_j and x_k .

3.1 Euclidean Distance

The Euclidean distance is most commonly used distance measure. It is also known as L_2 norm. The Euclidean distance, D_e, between two data points x_i and x_j is defined as:

$$D_e(x_i, x_j) = \left(\sum_{l=1}^{d} |x_{il} - x_{jl}|^2 \right)^{\frac{1}{2}} \tag{5}$$

where x_{il} and x_{jl} represent the l^{th} dimension of x_i and x_j respectively. It tends to form hyperspherical clusters. It satisfies all the above mentioned four conditions and therefore is a metric [21]. The strength of this measure is that clusters formed are invariant to translation and rotation in the feature space. This measure has disadvantages also. If one of the input attributes has a relatively large range, then it can overcome the other attributes [21].

3.2 Standardized Euclidean Distance

The standardized Euclidean distance is defined as the Euclidean distance between the data points divided by their standard deviation. The squared standardized Euclidean distance between x_i and x_j is mathematically described as:

$$D_{Se}(x_i, x_j) = (x_i - x_j) D^{-1} (x_i - x_j)^T \tag{6}$$

where D is the diagonal matrix with diagonal elements are given by va_j^2, which represents the variance of variable x_j over N data points. This measure is a metric as it satisfies the conditions of metric. When squared standardized Euclidean distance is multiplied by the geometric mean of the variances, it produces a diagonal Mahalanobis distance measure. The diagonal Mahalanobis distance fails to use the information of the diagonal in the covariance matrix [16].

3.3 Manhattan Distance

Manhattan distance between two data points is defined as the sum of the absolute differences of their coordinates. It is also known as a city block, rectilinear, taxicab or L_1 distance. It is mathematically defined as:

$$D_{Mn}(x_i, x_j) = \sum_{l=1}^{d} |x_{il} - x_{jl}| \tag{7}$$

The clusters formed using Manhattan distance tend to form rectangular shaped clusters. When all the features of dataset are binary in nature, the Manhattan distance acts as a Hamming distance [7]. It is also a metric. The advantage of over Euclidean distance is the reduced computation time [7]. Further it does not depend upon the translation and reflection of the coordinate system. The one disadvantage is that it depends upon the rotation of the coordinate system.

3.4 Mahalanbois Distance

Mahalanobis [12] introduced a new distance measure named as Mahalanobis distance. It is also known as quadratic distance. It is based on the correlations between variables by which different patterns can be identified and analyzed. The Mahalanobis distance is defined as:

$$D_{Ma}(x_i, x_j) = (x_i - x_j)V^{-1}(x_i - x_j)^T \qquad (8)$$

where V is covariance matrix. If the covariance matrix is identity matrix, the Mahalanobis distance reduces to the Euclidean distance. It differs from the Euclidean distance in that it takes into account the correlations of the dataset and is scale-invariant. It leads to violations of the triangle inequality and sensitive towards sampling fluctuations (Cherry et al., 1982). The Mahalanobis distance tends to form ellipsoidal clusters.

3.5 Cosine Distance

Cosine distance is a measure of dissimilarity between two vectors by measuring the cosine of the angle between them. It is defined as

$$D_{cos}(x_i, x_j) = 1 - \left(\frac{x_i^T x_j}{\|x_i\|\|x_j\|}\right) \qquad (9)$$

It is bounded between 0 and 1 if and are non-negative. It is used to measure cohesion within clusters [18]. This measure is not a distance metric and violates the triangle inequality. It is also invariant to scaling. It is unable to provide information on the magnitude of the differences. It is not invariant to shifts.

3.6 Correlation Distance

The correlation distance measure is derived from the Pearson correlation coefficient [8]. The correlation coefficient is used to measure the degree of linear dependency between two data points. The correlation based distance measure is mathematically formulated as:

$$D_{Corr}(x_i, x_j) = 1 - S_{CR}(x_i, x_j) \qquad (10)$$

$$S_{CR}(x_i, x_j) = \frac{\sum_{k=1}^{d}(m_{ik})(m_{jk})}{\sqrt{\sum_{k=1}^{d}(m_{ik})^2 \sum_{k=1}^{d}(m_{jk})^2}} \qquad (11)$$

where $m_{ik} = x_{ik} - \overline{x}_i$, $m_{jk} = x_{jk} - \overline{x}_j$, $\overline{x}_i = \frac{1}{d}\sum_{k=1}^{d} x_{ik}$ and $\overline{x}_j = \frac{1}{d}\sum_{k=1}^{d} x_{jk}$. This measure is not a distance metric. It tends to disclose the difference in shapes rather than to detect the magnitude of differences between two data points [21]. It is invariant to both scaling and translation.

3.7 Spearman Distance

The Spearman distance measure is derived from the Spearman correlation coefficient [5]. It can be defined as

$$D_{Spear}(x_i, x_j) = 1 - S_C(x_i, x_j) \qquad (12)$$

$$S_C(x_i, x_j) = \frac{\sum_{k=1}^{d}(m_{ik}^r)(m_{jk}^r)}{\sqrt{\sum_{k=1}^{d}(m_{ik}^r)^2 \sum_{k=1}^{d}(m_{jk}^r)^2}} \qquad (13)$$

where $m_{ik}^r = r(x_{ik}) - \overline{r}$, $m_{jk}^r = r(x_{jk}) - \overline{r}$. In Spearman rank correlation, each data value is replaced by their rank if the data in each vector is ordered by its value. Then Pearson correlation between the two rank vectors is computed instead of the data vectors. The Spearman rank correlation is an example of a non-parametric similarity measure. It is robust against outliers than the Pearson correlation. The disadvantage is that there is a loss of information when data are converted to ranks.

3.8 Chebyshev Distance

The Chebyshev distance calculates the maximum of the absolute differences between the features of a pair of data points. This distance is named after Panfnuty Chebyshev. It is also known as tchebyschev distance, maximum metric, chessboard distance, or metric. It is mathematically defined as

$$D_{Ch}(x_i, x_j) = max_{1 \leq l \leq d}(|x_{il} - x_{jl}|) \qquad (14)$$

This distance measure is a metric. The advantage is that it takes less time to decide the distances between data sets [15].

3.9 Canberra Distance

Lance and Williams [11] introduced a Canberra distance measure. It measures the sum of absolute fractional differences between the features of a pair of data points. It is mathematically defined as follows:

$$D_{Can}(x_i, x_j) = \sum_{l=1}^{d} \frac{|x_{il} - x_{jl}|}{|x_{il}| + |x_{jl}|} \qquad (15)$$

This distance measure is a metric. It is sensitive to a small change when both coordinates are near to zero.

3.10 Bray-Curtis Distance

The Bray-Curtis distance is also known as Sorensen distance [2]. This measure is computed using the absolute differences divided by the summation. It is defined as follows:

$$D_{Bc}(x_i, x_j) = \frac{\sum_{l=1}^{d} |x_{il} - x_{jl}|}{\sum_{l=1}^{d} (x_{il} + x_{jl})} \tag{16}$$

This distance measure is not a metric as it does not satisfy the triangle inequality property. The main drawback of this measure is that it is undefined if both data points are near zero values.

4 Experimental Results

This section provides a description of the datasets and demonstrate the efficiency of well-known clustering algorithms based on ten different distance measures. The results are evaluated and compared using some widely acceptable performance evaluation metrics such as accuracy, inter-cluster and intra-cluster distance [4]. Large value of accuracy measure is required for better clustering. Smaller value of intra-cluster and large value of inter-cluster distance is required for better clustering. All the results are evaluated in terms of 'mean' and 'standard deviation'. The standard deviation is used as a measure of robustness, which is shown in parenthesis.

4.1 Datasets used

Experiments are carried out with three artificial and six real-life datasets. A description of datasets is depicted in Table 1. The artificial datasets are named as Sp_4 _3, Sp_5 _2 and Sp_6 _2. These are taken from [1]. The six real-life datasets are obtained from UCI machine learning database [14].

Table 1: Description of Datasets Used

$DatasetName$	$Instances$	$Features$	$Classes$
Sp_5 _2	250	2	5
Sp_6 _2	300	2	6
Sp_4 _3	400	3	4
$Iris$	150	4	3
$Wine$	178	13	3
$Glass$	214	9	6
$Haberman$	306	3	2
$Bupa$	345	6	2
$Libras$	360	90	15

4.2 Parameter setting for the algorithms

The K-Means and K-Medoid were executed for 100 iterations. The parameters of the ACOC are as follows: evaporation rate = 0.1, number of ants = 20, and maximum number of iterations = 100 as mentioned in [17]. The parameters of the MHSC are as follows: harmony memory size = 15 and maximum number of iterations = 100. The pitch adjustment rate, harmony memory consideration rate and bandwidth are chosen as in [9, 10]. The value of K, number of clusters, for datasets equals the number of classes of the corresponding datasets as mentioned in Table 1.

4.3 Experimentation 1: Effect of distance measures on Hierarchical Techniques

Tables 2-10 show the effect of distance measures on accuracy for Sp_5 _2, Sp_6 _2, Sp_4 _3, $Iris$, $Wine$, $Glass$, $Haberman$, $Bupa$ and $Libras$ datasets respectively. The results reported in tables are the average values obtained over ten runs of algorithms. Figures 1-2 show the effect of distance measures on inter and intra-cluster distance.

Figure 1: Effect of distance measures on Inter-cluster distance for hierarchical techniques; (a) Sp_5 _2 (b) Sp_6 _2 (c) Sp_4 _3 (d) $Iris$ (e) $Wine$ (f) $Glass$ (g) $Glass$ (h) $Haberman$ (i) $Bupa.$

Figure 2: Effect of distance measures on Intra-cluster distance for hierarchical techniques; (a) Sp_5 _2 (b) Sp_6 _2 (c) Sp_4 _3 (d) $Iris$ (e) $Wine$ (f) $Glass$ (g) $Glass$ (h) $Haberman$ (i) $Bupa.$

For Sph_5 _2 (Table 2), it is found that the single and weighted linkage with Euclidean distance provide better accuracy over other distance measures. However, complete and average linkage with Bray-Curtis distance attain high accuracy as compared to other distance measures. From Figure 1(a), it has been analyzed that the complete, average, and weighted linkage clustering techniques using correlation distance produce well separated clusters. The single linkage clustering with Chebychev distance offers superior cluster separation over other distance measures. The complete, average, and weighted linkage clustering algorithms with Spearman distance produce compact clusters over other distance measures. The single linkage with Bray-Curtis distance gives best cluster compactness (Figure 2(a)).

Table 2: Effect of distance measures on accuracy of cluster formed for Sph_5 _2 dataset for Hierachical Clustering Techniques

Dist.Meas.	*S.Lin.*	*C.Lin.*	*A.Lin.*	*W.Lin.*
Eucl.	**0.596** **(0.000)**	0.948 (0.000)	0.940 (0.000)	**0.956** **(0.000)**
S.Eucl.	0.588 (0.000)	0.848 (0.000)	0.948 (0.000)	0.948 (0.000)
Manh.	0.356 (0.000)	0.940 (0.000)	0.944 (0.000)	0.860 (0.000)
Mahal.	0.588 (0.000)	0.928 (0.000)	0.948 (0.000)	0.912 (0.000)
Cos.	0.464 (0.000)	0.488 (0.000)	0.508 (0.000)	0.556 (0.000)
Corr.	0.400 (0.000)	0.416 (0.000)	0.404 (0.000)	0.408 (0.000)
Spear.	0.400 (0.000)	0.400 (0.000)	0.400 (0.000)	0.400 (0.000)
Cheb.	0.420 (0.000)	0.892 (0.000)	0.932 (0.000)	0.896 (0.000)
Canb.	0.216 (0.000)	0.736 (0.000)	0.944 (0.000)	0.812 (0.000)
Bray	0.204 (0.000)	**0.964** **(0.000)**	**0.964** **(0.000)**	0.720 (0.000)

Table 3: Effect of distance measures on accuracy of cluster formed for Sph_6 _2 dataset for Hierarchical Clustering Techniques

Dist.Meas.	*S.Lin.*	*C.Lin.*	*A.Lin.*	*W.Lin.*
Eucl.	**1.000** **(0.000)**	**1.000** **(0.000)**	**1.000** **(0.000)**	**1.000** **(0.000)**
S.Eucl.	**1.000** **(0.000)**	**1.000** **(0.000)**	**1.000** **(0.000)**	**1.000** **(0.000)**
Manh.	**1.000** **(0.000)**	**1.000** **(0.000)**	**1.000** **(0.000)**	**1.000** **(0.000)**
Mahal.	**1.000** **(0.000)**	**1.000** **(0.000)**	**1.000** **(0.000)**	**1.000** **(0.000)**
Cos.	0.483 (0.000)	0.560 (0.000)	0.577 (0.000)	0.577 (0.000)
Corr.	0.320 (0.000)	0.340 (0.000)	0.340 (0.000)	0.323 (0.000)
Spear.	0.320 (0.000)	0.320 (0.000)	0.320 (0.000)	0.320 (0.000)
Cheb.	**1.000** **(0.000)**	**1.000** **(0.000)**	**1.000** **(0.000)**	**1.000** **(0.000)**
Canb.	0.827 (0.000)	0.760 (0.000)	0.810 (0.000)	0.760 (0.000)
Bray	0.813 (0.000)	0.760 (0.000)	0.760 (0.000)	0.760 (0.000)

For Sph_6 _2 dataset (Table 3), all above-mentioned hierarchical clustering techniques provide well-separated and compact clusters with 100 percent accuracy using five distance measures as Euclidean, Standard Euclidean, Manhattan, Mahalanobis, and Chebychev.

Table 4: Effect of distance measures on accuracy of cluster formed for Sph_4 _3 dataset for Hierarchical Clustering Techniques

Dist.Meas.	*S.Lin.*	*C.Lin.*	*A.Lin.*	*W.Lin.*
Eucl.	**1.000** **(0.000)**	**1.000** **(0.000)**	**1.000** **(0.000)**	**1.000** **(0.000)**
S.Eucl.	**1.000** **(0.000)**	**1.000** **(0.000)**	**1.000** **(0.000)**	**1.000** **(0.000)**
Manh.	**1.000** **(0.000)**	**1.000** **(0.000)**	**1.000** **(0.000)**	**1.000** **(0.000)**
Mahal.	0.258 (0.000)	0.520 (0.000)	0.508 (0.000)	0.398 (0.000)
Cos.	0.358 (0.000)	0.323 (0.000)	0.338 (0.000)	0.340 (0.000)
Corr.	0.300 (0.000)	0.310 (0.000)	0.308 (0.000)	0.290 (0.000)
Spear.	0.302 (0.000)	0.302 (0.000)	0.302 (0.000)	0.302 (0.000)
Cheb.	**1.000** **(0.000)**	**1.000** **(0.000)**	**1.000** **(0.000)**	**1.000** **(0.000)**
Canb.	0.495 (0.000)	0.358 (0.000)	0.385 (0.000)	0.375 (0.000)
Bray	0.748 (0.000)	0.335 (0.000)	0.330 (0.000)	0.345 (0.000)

From Table 4, Figures 1(c) and 2(c), it is observed that all above-mentioned hierarchical clustering techniques provide well-separated and compact clusters with 100 percent accuracy using four distance measures (Euclidean, Standard Euclidean, Manhattan, and Chebychev) for Sph_4 _3 dataset.

Table 5: Effect of distance measures on accuracy of cluster formed for $Iris$ dataset for Hierarchical Clustering Techniques

Dist.Meas.	*S.Lin.*	*C.Lin.*	*A.Lin.*	*W.Lin.*
Eucl.	**0.680** **(0.000)**	0.840 (0.000)	0.907 (0.000)	0.900 (0.000)
S.Eucl.	0.660 (0.000)	0.787 (0.000)	0.687 (0.000)	0.567 (0.000)
Manh.	0.673 (0.000)	0.893 (0.000)	0.900 (0.000)	0.953 (0.000)
Mahal.	0.353 (0.000)	0.413 (0.000)	0.347 (0.000)	0.607 (0.000)
Cos.	0.660 (0.000)	0.840 (0.000)	0.660 (0.000)	**0.960** **(0.000)**
Corr.	0.660 (0.000)	0.853 (0.000)	**0.947** **(0.000)**	0.687 (0.000)
Spear.	0.673 (0.000)	0.673 (0.000)	0.673 (0.000)	0.673 (0.000)
Cheb.	**0.680** **(0.000)**	0.813 (0.000)	0.733 (0.000)	0.740 (0.000)
Canb.	0.627 (0.000)	**0.960** **(0.000)**	0.627 (0.000)	0.687 (0.000)
Bray	0.660 (0.000)	0.893 (0.000)	0.693 (0.000)	0.827 (0.000)

For $Iris$ dataset (Table 5), the single linkage with Euclidean or Chebyshev distance attains better accuracy than the other distance measures. However, it produces well-separated clusters with Mahalanobis distance (Figure 1(d)). From Figure 2(d), it has been found that single linkage gives compact clusters with four distance measures (Standard Euclidean, Cosine, Correlation, and Bray-Curtis). The complete linkage with Canberra distance provides higher accuracy than the other distances. It offers best cluster separation and compactness with Spearman distance (Figures 1(d) and 2(d)). The average linkage clustering with Correlation distance attains best accuracy. However, it gives well-separated clusters with Mahalanobis and compact clusters with Cosine distance. The weighted linkage using Cosine distance offeres best accuracy among other distance measures. It provides compact and best cluster separation with Spearman distance.

Table 6: Effect of distance measures on accuracy of cluster formed for $Wine$ dataset for Hierarchical Clustering Techniques

D.Meas.	*S.Lin.*	*C.Lin.*	*A.Lin.*	*W.Lin.*
Eucl.	**0.427** **(0.000)**	0.674 (0.000)	0.612 (0.000)	0.562 (0.000)
S.Eucl.	0.376 (0.000)	**0.837** **(0.000)**	0.388 (0.000)	0.618 (0.000)
Manh.	0.399 (0.000)	0.674 (0.000)	0.545 (0.000)	0.635 (0.000)
Mahal.	0.388 (0.000)	0.371 (0.000)	0.388 (0.000)	0.371 (0.000)
Cos.	0.410 (0.000)	0.562 (0.000)	0.448 (0.000)	0.483 (0.000)
Corr.	0.388 (0.000)	0.685 (0.000)	0.472 (0.000)	0.545 (0.000)
Spear.	0.387 (0.000)	0.612 (0.000)	0.612 (0.000)	0.589 (0.000)
Cheb.	**0.427** **(0.000)**	0.657 (0.000)	0.612 (0.000)	0.545 (0.000)
Canb.	0.387 (0.000)	0.652 (0.000)	0.646 (0.000)	**0.646** **(0.000)**
Bray	0.399 (0.000)	0.719 (0.000)	**0.725** **(0.000)**	0.573 (0.000)

For $Wine$ dataset, the single linkage with Euclidean or Chebychev distance produces well-separated clusters having optimal accuracy. It gives compact clusters with Bray-Curtis or City-Block distance (Figure 2(e)). The complete linkage with Standard Euclidean distance attains higher accuracy as compared to other distance measures. It produces well-separated clusters with Correlation distance (Figure 1(e)). The single and complete linkage produce compact clusters with Bray-Curtis distance. The average linkage with Bray-Curtis attains higher accuracy. The average linkage provides well-separated cluster with Euclidean or Chebychev distance and compact clusters with Mahalanobis distance. The weighted linkage with Canberra distance provides better accuracy than the other distance measures. It produces compact and well-separated clusters with City-Block distance.

Table 7: Effect of distance measures on accuracy of cluster formed for $Glass$ dataset for Hierarchical Clustering Techniques

D.Meas.	*S.Lin.*	*C.Lin.*	*A.Lin.*	*W.Lin.*
Eucl.	0.365	0.486	0.379	0.388
S.Eucl.	0.365	0.407	0.379	0.374
Manh.	**0.369** **(0.000)**	0.491 (0.000)	0.374 (0.000)	0.388 (0.000)
Mahal.	**0.369** **(0.000)**	0.421 (0.000)	0.374 (0.000)	0.379 (0.000)
Cos.	**0.369** **(0.000)**	**0.514** **(0.000)**	0.477 (0.000)	0.397 (0.000)
Corr.	**0.369** **(0.000)**	**0.514** **(0.000)**	**0.500** **(0.000)**	0.407 (0.000)
Spear.	0.365 (0.000)	0.486 (0.000)	0.477 (0.000)	0.477 (0.000)
Cheb.	0.365 (0.000)	0.495 (0.000)	0.477 (0.000)	**0.500** **(0.000)**
Canb.	0.346 (0.000)	0.463 (0.000)	0.365 (0.000)	0.365 (0.000)
Bray	**0.369** **(0.000)**	0.491 (0.000)	0.374 (0.000)	0.388 (0.000)

For $Glass$ dataset, the single linkage clustering provides higher accuracy over five distances (Manhattan, Mahalanobis, Cosine, Correlation and Bray-Curtis). The complete linkage attains good accuracy with Cosine and Correlation distances. The average linkage with Correlation distance produces superior accuracy. The weighted linkage provides better accuracy using Chebyshev distance. The single, complete, and weighted linkage with Standard Euclidean distance gives well-separated clusters (Figure 1(f)). The average linkage provides well-separated clusters using City-Block or Bray-Curtis distance. The single linkage with Canberra distance generates compact clusters. The complete and weighted linkage with Standard Euclidean distance gives clusters with good compactness (Figure 2(f)). The average linkage gives better compact clusters with Mahalanobis distance.

For $Haberman$ dataset (Table 8), the single linkage produces similar accuracy, inter-cluster distance and intra-cluster distance for all ten distance measures. Complete linkage attains best accuracy with Standard Euclidean distance. The average linkage with three distances (i.e. Euclidean, Spearman and Correlation) provide good accuracy. The complete and average linkage give well-separated and compact clusters with Spearman or Correlation distance (Figures 1(g) and 2(g)). The weighted linkage with five distance (Euclidean, Mahalanobis, Spearman, Correlation and Bray-Curtis)

Table 8: Effect of distance measures on accuracy of cluster formed for $Haberman$ dataset for Hierarchical Clustering Techniques

$D.Meas.$	$S.Lin.$	$C.Lin.$	$A.Lin.$	$W.Lin.$
$Eucl.$	**0.739** **(0.000)**	0.556 (0.000)	**0.739** **(0.000)**	**0.739** **(0.000)**
$S.Eucl.$	**0.739** **(0.000)**	**0.748** **(0.000)**	0.735 (0.000)	0.735 (0.000)
$Manh.$	**0.739** **(0.000)**	0.742 (0.000)	0.735 (0.000)	0.627 (0.000)
$Mahal.$	**0.739** **(0.000)**	0.745 (0.000)	0.735 (0.000)	**0.739** **(0.000)**
$Cos.$	**0.739** **(0.000)**	0.732 (0.000)	0.732 (0.000)	0.732 (0.000)
$Corr.$	**0.739** **(0.000)**	0.739 (0.000)	**0.739** **(0.000)**	**0.739** **(0.000)**
$Spear.$	**0.739** **(0.000)**	0.739 (0.000)	**0.739** **(0.000)**	**0.739** **(0.000)**
$Cheb.$	**0.739** **(0.000)**	0.552 (0.000)	0.735 (0.000)	0.637 (0.000)
$Canb.$	**0.739** **(0.000)**	0.569 (0.000)	0.582 (0.000)	0.582 (0.000)
$Bray$	**0.739** **(0.000)**	0.732 (0.000)	0.735 (0.000)	**0.739** **(0.000)**

provides similar accuracy. It produces compact clusters with Canberra distance and well-separated clusters with two distances (Spearman and Correlation).

Table 9: Effect of distance measures on accuracy of cluster formed for $Bupa$ dataset for Hierarchical Clustering Techniques

$D.Meas.$	$S.Lin.$	$C.Lin.$	$A.Lin.$	$W.Lin.$
$Eucl.$	**0.577** **(0.000)**	**0.577** **(0.000)**	0.557 (0.000)	**0.577** **(0.000)**
$S.Eucl.$	**0.577** **(0.000)**	0.559 (0.000)	0.571 (0.000)	**0.577** **(0.000)**
$Manh.$	0.571 (0.000)	0.574 (0.000)	0.562 (0.000)	0.557 (0.000)
$Mahal.$	**0.577** **(0.000)**	0.545 (0.000)	**0.577** **(0.000)**	**0.577** **(0.000)**
$Cos.$	**0.577** **(0.000)**	0.551 (0.000)	0.562 (0.000)	0.565 (0.000)
$Corr.$	**0.577** **(0.000)**	0.551 (0.000)	0.574 (0.000)	0.551 (0.000)
$Spear.$	**0.577** **(0.000)**	0.507 (0.000)	0.571 (0.000)	0.571 (0.000)
$Cheb.$	**0.577** **(0.000)**	0.554 (0.000)	0.557 (0.000)	0.571 (0.000)
$Canb.$	**0.577** **(0.000)**	0.522 (0.000)	0.562 (0.000)	0.554 (0.000)
$Bray$	**0.577** **(0.000)**	0.559 (0.000)	0.565 (0.000)	0.554 (0.000)

The results obtained for the $Bupa$ dataset (Table 9) show that the single linkage clustering produces similar accuracy for nine distance measures except Manhattan. It produces well-separated clusters using Mahalanobis distance and compact clusters using three distances named as Euclidean, Bray-Curtis, and Chebyshev (Figures 1(h) and 2(h)). The complete linkage using Euclidean distance produces accurate and compact clusters. It provides well-separated clusters with Bray-Curtis. The average linkage using Mahalanobis distance produces accurate and compact clusters. It provides well-separated clusters with City-Block. The weighted linkage attains best accuracy value on Euclidean, Standard Euclidean, and Mahalanobis distances. It generates compact clusters with Mahalanobis distance and well-separated clusters with Cosine distance.

Table 10: Effect of distance measures on accuracy of cluster formed for $Libras$ dataset for Hierarchical Clustering Techniques

$D.Meas.$	$S.Lin.$	$C.Lin.$	$A.Lin.$	$W.Lin.$
$Eucl.$	0.072 (0.000)	0.253 (0.000)	0.244 (0.000)	0.183 (0.000)
$S.Eucl.$	0.075 (0.000)	0.158 (0.000)	0.078 (0.000)	0.256 (0.000)
$Manh.$	0.078 (0.000)	0.147 (0.000)	**0.286** **(0.000)**	**0.267** **(0.000)**
$Mahal.$	0.069 (0.000)	0.075 (0.000)	0.069 (0.000)	0.069 (0.000)
$Cos.$	0.078 (0.000)	0.119 (0.000)	0.189 (0.000)	0.100 (0.000)
$Corr.$	**0.106** **(0.000)**	0.094 (0.000)	0.164 (0.000)	0.108 (0.000)
$Spear.$	0.069 (0.000)	**0.275** **(0.000)**	0.256 (0.000)	0.266 (0.000)
$Cheb.$	0.067 (0.000)	0.261 (0.000)	0.158 (0.000)	0.197 (0.000)
$Canb.$	0.072 (0.000)	0.068 (0.000)	0.068 (0.000)	0.150 (0.000)
$Bray$	0.072 (0.000)	0.244 (0.000)	0.139 (0.000)	0.128 (0.000)

For $Libras$ dataset results given in Table 10, show that the single linkage clustering using correlation distance provide good accuracy. The single linkage gives well-separated clusters using Cosine. The Complete linkage attains high accuracy over Spearman distance. The average and weighted linkage clustering with Manhattan distance gives better accuracy than the other distances. The complete and average linkage techniques produce well-separated clusters with Bray-Curtis distance (Figure 1(i)). The weighted linkage technique with Chebyshev distance generates separated clusters. Form Figure 2(i), it has been found that Mahalanobis distance provides compact clusters for all hierarchical techniques.

The aforementioned results indicate that the different distance measures with clustering techniques show different cluster quality value. The summarized results for hierarchical clustering techniques in terms of $accuracy$, $Inter - clusterDistance$, and $Intra - clusterDistance$ are tabulated in Tables 11, 12 and 13.

Table 11: Best distance measures corresponding to datasets and hierarchical clustering techniques in terms of Accuracy

Dataset	S.Lin.	C.Lin.	A.Lin.	W.Lin.
Sp_5 _2	Eucl.	Bray	Bray	Eucl.
Sp_6 _2	S5	S5	S5	S5
Sp_4 _3	Eucl. S.Eucl. Mahal. Cheb.	Eucl. S.Eucl. Mahal. Cheb.	Eucl. S.Eucl. Mahal. Cheb.	Eucl. S.Eucl. Mahal. Cheb.
Iris	Eucl. Cheb.	Canb.	Corr.	Cos.
Wine	Eucl. Cheb.	S.Eucl.	Bray	Canb.
Glass	Mahal. Manh. Corr. Bray Cos.	Cos. Corr.	Corr.	Cheb.
Haber.	All	S.Eucl.	Corr. Spear Eucl.	Spear Bray Mahal. Eucl. Corr.
Bupa	All except Manh.	Eucl.	Mahal.	Eucl. S.Eucl. Mahal.
Libras	Corr.	Spear	Manh.	Manh.
CMC	Corr.	Canb.	Cheb.	Corr.

Table 12: Best distance measures corresponding to datasets and hierarchical clustering techniques in terms of Inter-cluster Distance

Dataset	S.Lin.	C.Lin.	A.Lin.	W.Lin.
Sp_5 _2	Cheb	Corr.	Spear	Corr.
Sp_6 _2	S5	S5	S5	S5
Sp_4 _3	Mahal.	Eucl. S.Eucl. Mahal. Cheb.	Eucl. S.Eucl. Mahal. Cheb.	Eucl. S.Eucl. Mahal. Cheb.
Iris	Mahal.	Spear	Mahal. Cheb.	Spear
Wine	Eucl. Cheb.	Canb.	Eucl. Cheb.	Manh.
Glass	S.Eucl.	S.Eucl.	Manh. Bray	S.Eucl.
Haber	All	Corr. Spear	Corr. Spear	Corr. Spear
Bupa	Mahal.	Bray	Manh.	Cos.
Libras	Cos.	Bray	Bray	Cheb.

Table 13: Best distance measures corresponding to datasets and hierarchical clustering techniques in terms of Intra-cluster Distance

Dataset	S.Lin.	C.Lin.	A.Lin.	W.Lin.
Sp_5 _2	Bray	Spear	Spear	Spear
Sp_6 _2	S5	S5	S5	S5
Sp_4 _3	Eucl. S.Eucl. Mahal. Cheb.	Eucl. S.Eucl. Mahal. Cheb.	Eucl. S.Eucl. Mahal. Cheb.	Eucl. S.Eucl. Mahal. Cheb.
Iris	S.Eucl. Cos. Corr.	Spear	Cos.	Spear
Wine	Manh. Bray	Bray	Mahal.	Manh.
Glass	Canb.	S.Eucl.	Mahal.	S.Eucl.
Haber.	All	Corr. Spear	Corr. Spear	Canb.
Bupa	Eucl. Bray Cheb.	Eucl.	Mahal.	Mahal.
Libras	Mahal.	Mahal.	Mahal.	Mahal.

4.4 Experimentation 2: Effect of distance measures on Partitional Techniques

Tables 14-22 show the effect of distance measures on accuracy for $Sp\_5\ \_2$, $Sp\_6\ \_2$, $Sp\_4\ \_3$, $Iris$, $Wine$, $Glass$, $Haberman$, $Bupa$ and $Libras$ datasets respectively. The results reported in tables are the average values obtained over ten runs of algorithms. Figures 3-4 show the effect of distance measures on inter and intra-cluster distance.

Figure 3: Effect of distance measures on Inter-cluster Distance for partitional techniques; (a) $Sp\_5\ \_2$ (b) $Sp\_6\ \_2$ (c) $Sp\_4\ \_3$ (d) $Iris$ (e) $Wine$ (f) $Glass$ (g) $Glass$ (h) $Haberman$ (i) $Bupa$.

Figure 4: Effect of distance measures on Intra-cluster Distance for partitional techniques; (a) $Sp\_5\ \_2$ (b) $Sp\_6\ \_2$ (c) $Sp\_4\ \_3$ (d) $Iris$ (e) $Wine$ (f) $Glass$ (g) $Glass$ (h) $Haberman$ (i) $Bupa$.

For $Sph\_5\ \_2$ dataset (Table 14), MHSC and K-Medoid attain best accuracy with Euclidean distance. The K-Means clustering algorithm with Chebychev distance attains best accuracy. The ACOC algorithm provides higher accuracy with Bray-Curtis distance. The K-Medoid and ACOC with Spearman distance produce compact and well-separated clusters (Figure 3(a)). The K-Means with Spearman distance gives well separated clusters. The MHSC with Canberra distance offers superior cluster separation over other distance

Table 14: Effect of distance measures on accuracy of cluster formed for Sph_5 _2 dataset for Partitional Clustering Techniques

$Dist.Meas.$	KM	KMD	$ACOC$	$MHSC$
$Eucl.$	0.958 (0.0144)	**0.907** **(0.137)**	0.940 (0.013)	**0.956** **(0.074)**
$S.Eucl.$	0.965 (0.018)	0.867 (0.146)	0.948 (0.011)	0.948 (0.078)
$Manh.$	0.863 (0.101)	0.815 (0.086)	0.944 (0.008)	0.860 (0.086)
$Mahal.$	0.959 (0.012)	0.895 (0.117)	0.948 (0.008)	0.912 (0.095)
$Cos.$	0.583 (0.039)	0.529 (0.035)	0.508 (0.006)	0.556 (0.104)
$Corr.$	0.400 (0.000)	0.409 (0.017)	0.404 (0.004)	0.408 (0.107)
$Spear.$	0.400 (0.000)	0.400 (0.000)	0.400 (0.011)	0.400 (0.066)
$Cheb.$	**0.978** **(0.006)**	0.855 (0.140)	0.932 (0.009)	0.896 (0.048)
$Canb.$	0.943 (0.021)	0.851 (0.068)	0.944 (0.011)	0.812 (0.050)
$Bray$	0.887 (0.107)	0.895 (0.062)	**0.964** **(0.006)**	0.720 (0.011)

measures. The MHSC gives compact clusters with Euclidean distance (Figure 4(a)).

Table 15: Effect of distance measures on accuracy of cluster formed for Sph_6 _2 dataset for Partitional Clustering Techniques

$Dist.Meas.$	KM	KMD	$ACOC$	$MHSC$
$Eucl.$	0.794 (0.143)	**0.941** **(0.109)**	**0.213** **(0.008)**	0.804 (0.056)
$S.Eucl.$	0.739 (0.135)	0.803 (0.136)	0.209 (0.007)	0.803 (0.076)
$Manh.$	0.792 (0.146)	0.873 (0.137)	0.209 (0.005)	0.829 (0.007)
$Mahal.$	0.799 (0.139)	0.915 (0.118)	0.205 (0.007)	0.803 (0.039)
$Cos.$	0.578 (0.054)	0.546 (0.054)	0.202 (0.016)	0.768 (0.083)
$Corr.$	0.325 (0.012)	0.300 (0.014)	0.205 (0.012)	0.648 (0.090)
$Spear.$	0.333 (0.000)	0.333 (0.000)	0.204 (0.008)	0.678 (0.079)
$Cheb.$	0.850 (0.124)	0.787 (0.202)	0.204 (0.009)	0.790 (0.056)
$Canb.$	0.851 (0.123)	0.825 (0.154)	0.201 (0.006)	0.890 (0.116)
$Bray$	**0.885** **(0.122)**	0.793 (0.187)	0.199 (0.005)	**0.991** **(0.012)**

For Sph_6 _2 dataset (Table 15), K-Means produces compact and accurate clusters with Bray-Curtis distance. A careful look at Figure 3(b) reveals that K-Means with Canberra distance provide well-separated clusters. From Figures 3(b) and 4(b), it has been seen that K-Medoid with Euclidean distance gives well-separated and compact clusters having accuracy higher than the other distance measures. The ACOC technique with Euclidean distance attains better accuracy. However, it produces well-separated and compact clusters with Mahalanobis and Cosine distance respectively. The MHSC with Bray-Curtis produces well-separated and compact clusters having higher accuracy when compared with other measures.

Table 16: Effect of distance measures on accuracy of cluster formed for Sph_4 _3 dataset for Partitional Clustering Techniques

$Dist.Meas.$	KM	KMD	$ACOC$	$MHSC$
$Eucl.$	0.909 (0.168)	**0.957** **(0.121)**	0.286 (0.008)	**0.973** **(0.077)**
$S.Eucl.$	0.825 (0.187)	0.828 (0.184)	0.286 (0.008)	0.950 (0.091)
$Manh.$	0.382 (0.025)	0.778 (0.184)	0.279 (0.007)	0.871 (0.146)
$Mahal.$	0.520 (0.022)	0.526 (0.018)	0.278 (0.011)	0.771 (0.143)
$Cos.$	0.413 (0.042)	0.415 (0.055)	0.270 (0.008)	0.702 (0.139)
$Corr.$	0.301 (0.004)	0.295 (0.004)	0.269 (0.006)	0.710 (0.109)
$Spear.$	0.293 (0.000)	0.295 (0.000)	0.272 (0.006)	0.657 (0.164)
$Cheb.$	**1.000** **(0.000)**	0.826 (0.186)	0.285 (0.007)	0.892 (0.156)
$Canb.$	0.652 (0.217)	0.744 (0.208)	0.283 (0.009)	0.827 (0.179)
$Bray$	0.604 (0.008)	0.818 (0.077)	**0.288** **(0.009)**	0.656 (0.081)

For Sph_4 _3 dataset (Table 16), K-Means with Chebychev distance offers highest accuracy over other distance measures. The K-Means with Chebychev distance generates well-separated and compact clusters. The K-Medoid provides better accuracy with compact clusters using Euclidean distance. While, it generates well-separated clusters using Bray-Curtis distance (Figure 3(c)). The ACOC technique with Bray-Curtis distance generates accurate, compact and well-separated clusters. The MHSC with Euclidean distance produces accurate, well-separated, and compact clusters as compared to other distance measures.

For Iris dataset, the K-Means clustering algorithm provides better accuracy with Chebychev distance. The K-Medoid technique with Manhattan distance attains best accuracy. Both K-Means and K-Medoid produce compact and well-separated clusters with Spearman distance. The ACOC with Spearman distance attains better accuracy. It produces compact clusters with

Table 17: Effect of distance measures on accuracy of cluster formed for $Iris$ dataset for Partitional Clustering Techniques

$Dist.Meas.$	KM	KMD	$ACOC$	$MHSC$
$Eucl.$	0.844	0.718	0.396	0.867
	(0.131)	(0.190)	(0.013)	(0.047)
$S.Eucl.$	0.844	0.713	0.392	0.882
	(0.131)	(0.196)	(0.009)	(0.068)
$Manh.$	0.767	**0.841**	0.395	0.885
	(0.166)	**(0.133)**	(0.009)	(0.055)
$Mahal.$	0.778	0.526	0.389	0.868
	(0.036)	(0.161)	(0.021)	(0.061)
$Cos.$	0.803	0.828	0.382	0.848
	(0.235)	(0.167)	(0.011)	(0.138)
$Corr.$	0.801	0.839	0.385	**0.891**
	(0.221)	(0.176)	(0.008)	**(0.072)**
$Spear.$	0.667	0.667	**0.415**	0.716
	(0.000)	(0.000)	**(0.033)**	(0.132)
$Cheb.$	**0.887**	0.768	0.385	0.871
	(0.000)	(0.174)	(0.009)	(0.032)
$Canb.$	0.867	0.774	0.366	0.883
	(0.148)	(0.191)	(0.005)	(0.067)
$Bray$	0.798	0.798	0.364	0.862
	(0.165)	(0.180)	(0.007)	(0.066)

Bray-Curtis and well-separated clusters with Canberra distance (Figures 3(d) and 4(d)). The MHSC technique provides well-separated and accurate clusters using Correlation distance. While, it offers compact clusters with Euclidean distance.

Table 18: Effect of distance measures on accuracy of cluster formed for $Wine$ dataset for Partitional Clustering Techniques

$D.Meas.$	KM	KMD	$ACOC$	$MHSC$
$Eucl.$	0.669	0.631	**0.432**	**0.691**
	(0.059)	(0.080)	**(0.021)**	**(0.054)**
$S.Eucl.$	0.669	0.704	0.399	0.671
	(0.059)	(0.008)	(0.015)	(0.061)
$Manh.$	0.669	0.650	0.422	0.663
	(0.066)	(0.078)	(0.028)	(0.060)
$Mahal.$	0.609	0.459	0.368	0.638
	(0.119)	(0.036)	(0.066)	(0.059)
$Cos.$	0.685	0.657	0.409	0.689
	(0.018)	(0.000)	(0.021)	(0.048)
$Corr.$	0.685	0.669	0.415	0.665
	(0.012)	(0.013)	(0.012)	(0.058)
$Spear.$	0.664	0.608	0.389	0.655
	(0.034)	(0.088)	(0.023)	(0.073)
$Cheb.$	0.652	0.684	0.419	0.678
	(0.069)	(0.054)	(0.021)	(0.056)
$Canb.$	**0.891**	0.621	0.413	0.689
	(0.129)	(0.077)	(0.008)	(0.042)
$Bray$	0.717	**0.719**	0.419	0.678
	(0.003)	**(0.007)**	(0.008)	(0.072)

For Wine dataset, the K-Means technique attains higher accuracy with Canberra distance. It produces well-separated clusters with Chebychev distance and compact clusters with Bray-Curtis distance (Figures 3(e) and 4(e)). K-Medoid attains optimal accuracy using Bray-Curtis distance. It gives well-separated clusters with Spearman distance and compact clusters with Mahalanobis distance. The ACOC technique with Euclidean distance provides better accuracy than the other measures. It provides compact clusters with Bray-Curtis and well-separated clusters with Canberra distance. The MHSC technique with Euclidean distance offers accurate and well-separated clusters. It produces compact clusters with Cosine distance.

Table 19: Effect of distance measures on accuracy of cluster formed for $Glass$ dataset for Partitional Clustering Techniques

$D.Meas.$	KM	KMD	$ACOC$	$MHSC$
$Eucl.$	0.508	0.463	0.329	0.494
	(0.036)	(0.053)	(0.019)	(0.036)
$S.Eucl.$	0.506	0.473	0.331	0.489
	(0.045)	(0.039)	(0.025)	(0.025)
$Manh.$	0.512	0.476	0.341	0.451
	(0.031)	(0.021)	(0.015)	(0.031)
$Mahal.$	0.415	0.398	0.244	0.485
	(0.039)	(0.024)	(0.017)	(0.064)
$Cos.$	0.517	0.474	0.337	**0.502**
	(0.034)	(0.028)	(0.017)	**(0.027)**
$Corr.$	**0.519**	**0.485**	0.342	0.501
	(0.024)	**(0.033)**	(0.015)	(0.032)
$Spear.$	0.481	0.457	0.339	0.494
	(0.035)	(0.014)	(0.007)	(0.034)
$Cheb.$	0.496	0.445	0.330	0.499
	(0.028)	(0.049)	(0.016)	(0.032)
$Canb.$	0.416	0.408	**0.360**	0.394
	(0.068)	(0.036)	**(0.014)**	(0.054)
$Bray$	**0.519**	**0.485**	0.345	0.476
	(0.029)	**(0.018)**	(0.013)	(0.042)

For Glass dataset, the K-Means and K-Medoid provides best accuracy with Correlation and Bray-Curtis distances. K-Means, K-Medoid, and ACOC techniques attain well-separated and compact clusters with Canberra distance (Figures 3(f) and 4(f)). ACOC technique provides good accuracy over Canberra distance. The MHSC technique gives accurate clusters with Cosine distance and well-separated clusters with Spearman distance. It produces compact clusters with Mahalanobis distance.

For Haberman dataset (Table 20), the K-Means using Spearman distance gives accurate and well-separated clusters. It generates compact clusters with Chebychev distance (Figure 4(g)). K-Medoid using Spearman distance produces accurate, compact, and well-separated clusters. ACOC attains best accuracy

Table 20: Effect of distance measures on accuracy of cluster formed for *Haberman* dataset for Partitional Clustering Techniques

D.Meas.	*KM*	*KMD*	*ACOC*	*MHSC*
Eucl.	0.509 (0.011)	0.594 (0.104)	0.672 (0.014)	0.554 (0.040)
S.Eucl.	0.514 (0.008)	0.547 (0.075)	0.561 (0.015)	0.542 (0.036)
Manh.	0.579 (0.107)	0.575 (0.094)	0.683 (0.014)	0.532 (0.024)
Mahal.	0.522 (0.020)	0.546 (0.085)	0.523 (0.015)	0.548 (0.033)
Cos.	0.513 (0.003)	0.536 (0.016)	0.713 (0.011)	**0.583** **(0.087)**
Corr.	0.509 (0.000)	0.531 (0.018)	0.711 (0.012)	0.558 (0.062)
Spear.	**0.647** **(0.000)**	**0.664** **(0.006)**	0.647 (0.014)	0.581 (0.078)
Cheb.	0.534 (0.011)	0.529 (0.029)	0.676 (0.011)	0.529 (0.024)
Canb.	0.527 (0.001)	0.529 (0.038)	0.699 (0.010)	0.575 (0.063)
Bray	0.509 (0.000)	0.550 (0.087)	**0.724** **(0.006)**	0.534 (0.027)

on Bray-Curtis distance. It generates well-separated clusters with Correlation and compact clusters with Spearman distance. The MHSC technique provides higher accuracy over cosine distance. It gives well-separated clusters with Spearman and compact clusters with City-Block distance (Figures 3(g) and 4(g)).

Table 21: Effect of distance measures on accuracy of cluster formed for *Bupa* dataset for Partitional Clustering Techniques

D.Meas.	*KM*	*KMD*	*ACOC*	*MHSC*
Eucl.	0.551 (0.000)	0.529 (0.004)	0.564 (0.018)	0.538 (0.019)
S.Eucl.	0.551 (0.001)	0.531 (0.002)	0.532 (0.018)	0.527 (0.022)
Manh.	0.548 (0.000)	0.536 (0.013)	0.567 (0.017)	0.544 (0.018)
Mahal.	0.534 (0.029)	0.544 (0.025)	0.518 (0.013)	0.547 (0.025)
Cos.	0.507 (0.000)	0.510 (0.000)	0.567 (0.011)	0.539 (0.020)
Corr.	0.536 (0.000)	0.530 (0.001)	0.575 (0.007)	0.536 (0.017)
Spear.	**0.622** **(0.002)**	**0.622** **(0.002)**	**0.579** **(0.009)**	**0.617** **(0.012)**
Cheb.	0.542 (0.000)	0.539 (0.031)	0.574 (0.011)	0.536 (0.033)
Canb.	0.530 (0.000)	0.545 (0.000)	0.578 (0.007)	0.552 (0.028)
Bray	0.510 (0.000)	0.522 (0.000)	0.576 (0.009)	0.539 (0.021)

For the Bupa dataset (Table 21), K-Means, K-Medoid, and ACOC produces accurate clusters with compactness using Spearman distance. K-Means gives well-separated clusters using Standard Euclidean distance. K-Medoid and ACOC generate well-separated clusters with Spearman and Bray-Curtis distances respectively. The MHSC algorithm provides better accuracy with Spearman distance. It gives compact clusters with Canberra and well-separated clusters with Cosine distance (Figures 3(h) and 4(h)).

Table 22: Effect of distance measures on accuracy of cluster formed for *Libras* dataset for Partitional Clustering Techniques

D.Meas.	*KM*	*KMD*	*ACOC*	*MHSC*
Eucl.	0.188 (0.048)	0.191 (0.048)	0.071 (0.007)	0.157 (0.044)
S.Eucl.	0.207 (0.065)	0.134 (0.094)	0.064 (0.013)	0.159 (0.048)
Manh.	0.175 (0.070)	0.151 (0.062)	**0.079** **(0.009)**	0.140 (0.058)
Mahal.	0.084 (0.028)	0.066 (0.009)	0.063 (0.004)	0.145 (0.039)
Cos.	0.235 (0.079)	**0.198** **(0.044)**	0.074 (0.011)	0.155 (0.049)
Corr.	**0.259** **(0.049)**	0.182 (0.047)	0.075 (0.007)	0.159 (0.062)
Spear.	0.224 (0.037)	0.117 (0.051)	0.078 (0.007)	0.152 (0.029)
Cheb.	0.248 (0.064)	0.186 (0.037)	0.078 (0.011)	0.121 (0.046)
Canb.	0.183 (0.052)	0.178 (0.060)	0.078 (0.007)	0.147 (0.052)
Bray	0.175 (0.059)	0.182 (0.094)	0.071 (0.009)	**0.193** **(0.047)**

For *Libras* dataset results given in Table 22, show that the K-Means using Correlation distance attains best accuracy. However, it gives compact clusters with Bray-Curtis and well-separated clusters with Canberra distance. K-Medoid provides accurate clusters on Cosine distance and compact clusters with Mahalanobis. It generates well-separated clusters with Euclidean distance. ACOC provides accurate clusters on Manhattan distance and compact clusters with Chebyshev. However, it generates well-separated clusters with Spearman distance. The MHSC technique attains high accuracy over Bray-Curtis distance. It gives well-separated and compact clusters with City-Block distance (Figures 3(i) and 4(i)).

The aforementioned results indicate that the different distance measures with clustering techniques show different cluster quality value. The summarized results for partitional techniques are tabulated in Tables 23, 24 and 25.

Table 23: Best distance measures corresponding to datasets and partitional clustering techniques in terms of Accuracy

Dataset	*KM*	*KMD*	*ACOC*	*MHSC*
Sp_5 _2	*Cheb.*	*Eucl.*	*Bray.*	*Eucl.*
Sp_6 _2	*Bray*	*Eucl.*	*Eucl.*	*Bray*
Sp_4 _3	*Eucl.*	*Eucl.*	*Bray*	*Eucl.*
Iris	*Cheb.*	*Mahal.*	*Spear*	*Corr.*
Wine	*Canb.*	*Bray*	*Eucl.*	*Eucl.*
Glass	*Corr.* *Bray*	*Corr.* *Bray*	*Canb.*	*Cos.*
Haber.	*Spear*	*Spear*	*Bray*	*Cos.*
Bupa	*Spear*	*Spear*	*Spear*	*Spear*
Libras	*Corr.*	*Cos.*	*Manh.*	*Bray*

Table 24: Best distance measures corresponding to datasets and partitional clustering techniques in terms of Inter-cluster Distance

Dataset	*KM*	*KMD*	*ACOC*	*MHSC*
Sp_5 _2	*Spear*	*Spear*	*Spear*	*Canb.*
Sp_6 _2	*Canb.*	*Eucl.*	*Mahal.*	*Bray*
Sp_4 _3	*Cheb.*	*Bray*	*Bray*	*Eucl.*
Iris	*Spear*	*Spear*	*Canb.*	*Corr.*
Wine	*Cheb.*	*Spear*	*Canb.*	*Eucl.*
Glass	*Canb.*	*Canb.*	*Canb.*	*Spear*
Haber.	*Spear*	*Spear*	*Corr.*	*Spear*
Bupa	*S.Eucl.*	*Spear*	*Bray*	*Cos.*
Libras	*Canb.*	*Eucl.*	*Spear*	*Manh.*

Table 25: Best distance measures corresponding to datasets and partitional clustering techniques in terms of Intra-cluster Distance

Dataset	*KM*	*KMD*	*ACOC*	*MHSC*
Sp_5 _2	*Corr.*	*Spear*	*Spear*	*Eucl.*
Sp_6 _2	*Bray*	*Eucl.*	*Cos.*	*Bray*
Sp_4 _3	*Cheb.*	*Eucl.*	*Bray*	*Eucl.*
Iris	*Spear*	*Spear*	*Bray*	*Eucl.*
Wine	*Bray*	*Mahal.*	*Bray*	*Cos.*
Glass	*Canb.*	*Canb.*	*Canb.*	*Mahal.*
Haber.	*Cheb.*	*Spear*	*Spear*	*Manh.*
Bupa	*Spear*	*Spear*	*Spear*	*Canb.*
Libras	*Bray*	*Mahal.*	*Cheb.*	*Manh.*

5 Conclusion

In this paper, performance of ten commonly used distance measures in clustering techniques has been evaluated. The eight well-known clustering algorithms are evaluated on ten different datasets. The experimental results are evaluated in terms of accuracy, inter-cluster and intra-cluster distances. It has been observed that there is no single best distance measure for all datasets, or for all quality measures. The appropriateness of a distance measure is dependent on nature of data and clustering technique. On basis of our experimentation, we have reported a set of suitable distance measures for a particular combination of distance and clustering techniques.

References

[1] Bandyopadhyay, S. and Maulik, U. Genetic clustering for automatic evolution of clusters and application to image classification. *Pattern Recognition*, 35(6):1197–1208, 2002.

[2] Bray, J. R. and Curtis, J. T. An ordination of the upland forest communities of southern wisconsin. *Ecological Monographs*, 27(4):325–349, 1957.

[3] Cao, F., Liang, J., Li, D., Bai, L., and Dang, C. A dissimilarty measure for the k-modes clustering algorithm. *Knowledge-Based Systems*, 26(1):120–127, 2012.

[4] Chen, J., Zhao, Z., Ye, J., and Liu, H. Nonlinear adaptive distance metric learning for clustering. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pages 123–132, 2007.

[5] Fulekar, M. H. *Bioinformatics: Applications in Life and Environmental Sciences*. Springer, 2009.

[6] Huang, A. Similarity measures for text document clustering. In *Proceedings of the Sixth New Zealand Computer Research Student Conference*, pages 49–56, 2008.

[7] Jain, A. K. and Dubes, R. C. *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, 1988.

[8] Kaufman, L. and Rousseeuw, P. J. *Finding groups in data: an introduction to cluster analysis*. John Wiley and Sons, New York, 1990.

[9] Kumar, V., Chhabra, J. K., and Kumar, D. Effect of harmony serach parameters' variation in clustering. *Procedia Technology*, 6:265–274, 2012.

[10] Kumar, V., Chhabra, J. K., and Kumar, D. Clustering using modified harmony search algorithm. *International Journal of Computational Intelligence Studies*, 3(2):113–133, 2014.

[11] Lance, G. N. and Williams, W. T. Computer programs for hierarchical polythetic classification (similarity analyses). *Computer*, 9(1):60–64, 1966.

[12] Mahalanbois, P. C. On the generalized distance in statistics. In *Proceedings of National Institute of Science*, pages 49–55, 1936.

[13] Maulik, U. and Bandyopadhyay, S. Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12):1650–1654, 2002.

[14] Newman, C. L., Blake, D., and Merz, C. J. Uci repository of machine learning databases, 1998.

[15] Potolea, R., Cacoveanu, S., and Lemnaru, C. Meta-learning framework for prediction strategy evaluation. In *Proceedings of International Conference on Enterprise Information Systems*, pages 280–295, Magdeburg, Germany, 2011.

[16] Prekopcsak, Z. and Lemire, D. Time series classification by class-specific mahalanobis distance measures. *Advances in Data Analysis and Classification*, 2(1):49–55, 2012.

[17] Shelokar, P. S., Jayaraman, V. K., and Kulkarni, B. D. An ant colony approach for clustering. *Analytica Chimica Acta*, 509(2):187–195, 2004.

[18] Tan, P. N., Steinbach, M., and Kumar, V. *Introduction to data mining*. Addison-Wesley, 2005.

[19] Vadapalli, S. on the ignored aspects of data clustering. In *GHC of Women in Computing*, pages 6–10, 2004.

[20] Vimal, A., Valluri, S., and Karlapalem, K. An experiment with distance measures for clustering. Technical Report IIIT/TR/2008/132, Center for Data Engineering, IIIT, Hyderabad, July 2008.

[21] Xu, R. and Wunsch, D. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):120–127, 2005.