

Impact of Distance Measures on the Performance of Clustering Algorithms

Vijay Kumar¹, Jitender Kumar Chhabra², Dinesh Kumar³

¹CSE Department, JCDDM College of Engineering, Sirsa, Haryana, INDIA,

²Computer Engineering Department, NIT, Kurukshetra, Haryana, India,

³CSE Department, GJUS&T, Hisar, Haryana, INDIA

vijaykumarchahar@gmail.com,jitenderchhabra@gmail.com,dinesh.chutani@yahoo.com

Abstract. Distance measure plays a vital role in clustering algorithms. Selecting the right distance measure for given dataset is a challenging problem. In this paper, the effect of six distance measures on three clustering algorithms, K-means, single linkage and average linkage is investigated. The distance measures include euclidean, euclidean squared, manhattan, mahalanbois, cosine similarity, and pearson correlation. We describe all the distance measures pointing out their strengths and weaknesses. The performance of clustering algorithms on distance measures are evaluated on two artificial and four real life datasets. Experimental results show the impact of distance measures when used for different datasets.

Keywords: Clustering, Distance Measure, Clustering Algorithms

1 Introduction

Clustering is a process of grouping data points based on the similarity/ dissimilarity between them measures. It has a wide variety of applications in many areas such as data mining, pattern recognition and data compression etc. In general, clustering techniques are classified into two main categories: hierarchical and partitional clustering. These clustering techniques are either directly or indirectly based on distance measures. The distance measures depend upon the applications. Different applications require different notations of similarity measure between data objects. Right selection of 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 [8].

The main contribution of this paper is to demonstrate the impact of distance measures on three well-known clustering algorithms and discuss their relative strengths and weaknesses. These are evaluated over artificial as well as real-life datasets. The remainder of this paper is organized as follows. In section 2, we review the clustering algorithms. Distance functions are summarized in section 3. The results of the experiments are presented and discussed in section 4. Finally, a concluding remark is drawn in section 5.

2 Clustering Algorithms

The clustering algorithms are used to partition dataset $X = \{x_1, \dots, x_j, \dots, x_N\}$, where $x_j = (x_{j1}, x_{j2}, \dots, x_{jd}) \in R^d$ into a number of clusters, say $K, (C_1, \dots, C_K)$. The partition matrix $U(X)$ is represented as $U = [u_{kj}]$, $k = 1, 2, \dots, K$, and $j = 1, 2, \dots, N$, where u_{kj} is membership of data point x_j to clusters C_k . The $u_{kj} = 1$ if $x_j \in C_k$; otherwise, $u_{kj} = 0$. The three clustering algorithms used in this paper are single linkage clustering, average linkage clustering, and K-means. The single linkage clustering computes the distances between two clusters p and q such as [4]

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

Whereas the average linkage clustering uses the average of pairwise distance between points in two clusters p and q as:

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

The third clustering algorithm is K-Means clustering algorithm. It seeks an optimal partition of data by minimizing the sum-of-squared-error criterion with an iterative optimization procedure such as [4]

$$J(U, V) = \sum_{j=1}^N \sum_{i=1}^K u_{ij} \|x_j - v_i\|^2 \quad (3)$$

where v_i is the center of cluster C_i .

3 Distance Measures

Most of the clustering algorithms are based on the distance measures between data points. The large value of distance measure represents greater the difference between objects. The following are the most commonly used distance measures [8].

3.1 Euclidean Distance

The euclidean distance, also known as L_2 norm, between two data point x_i and x_j is defined as:

$$D_e(x_i, x_j) = \left(\sum_{l=1}^d |x_{il} - x_{jl}|^2 \right)^{\frac{1}{2}} \quad (4)$$

where x_{il} and x_{jl} represent the l^{th} dimension of x_i and x_j respectively. It tends to form hyperspherical clusters. The strength of this measure is that clusters formed are invariant to translation and rotation in the feature space [3]. This measure has disadvantages also. If one of the input attributes has a relatively large range, then it can overcome the other attributes [9].

3.2 Standardized Euclidean Distance

The standardized Euclidean distance between x_i and x_j is mathematically described as:

$$D_{se}(x_i, x_j) = (x_i - x_j)D^{-1}(x_i - x_j)^T \quad (5)$$

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. 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 [7].

3.3 Manhattan Distance

Manhattan distance also known as a city-block, rectilinear or L_1 distance, is mathematically defined as:

$$D_{mn}(x_i, x_j) = \sum_{l=1}^d |x_{il} - x_{jl}| \quad (6)$$

The clusters formed using Manhattan distance tend to form rectangular shaped clusters. Its advantage over Euclidean distance is the reduced computation time [4]. 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 Mahalanobis Distance

Mahalanobis introduced a new distance measure named as Mahalanobis distance [6]. It is based on the correlations between variables by which different patterns can be identified and analyzed. The squared Mahalanobis is defined as follows:

$$D_{ma}(x_i, x_j) = (x_i - x_j)V^{-1}(x_i - x_j)^T \quad (7)$$

where V is covariance matrix. It differs from the Euclidean distance in that it takes into account the correlations of the dataset. The Mahalanobis distance tends to form ellipsoidal clusters.

3.5 Cosine Distance

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

$$D_{cos}(x_i, x_j) = 1 - \left(\frac{x_i^T x_j}{\|x_i\| \|x_j\|} \right) \quad (8)$$

It is used to measure cohesion within clusters [4]. It is also invariant to scaling. It is unable to provide information on the magnitude of the differences.

3.6 Correlation Distance

The correlation distance measure is derived from the Pearson correlation coefficient [5]. It is mathematically formulated as:

$$D_{corr}(x_i, x_j) = 1 - \frac{\sum_{k=1}^d (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^d (x_{ik} - \bar{x}_i)^2 \sum_{k=1}^d (x_{jk} - \bar{x}_j)^2}} \quad (9)$$

where $\bar{x}_i = \frac{1}{d} \sum_{k=1}^d x_{ik}$ and $\bar{x}_j = \frac{1}{d} \sum_{k=1}^d x_{jk}$.

4 Experimental Results and discussions

In this section, we demonstrate the efficiency of clustering algorithms based on six different distance measures. The results are evaluated and compared using some widely acceptable performance evaluation metrics such as accuracy[2], and inter-cluster distance. Large value of these measures are required for better clustering.

4.1 Datasets used

Six datasets are used for experiments. Two of them are artificial datasets (*Sph_5_2* and *Sph_6_2*) and four are real-life datasets (Iris, Wine, Bupa and CMC). These are taken from [1]. The real life datasets are obtained from the UCI machine learning database [10].

4.2 Evaluation on clustering efficiency

Tables 1-6 show the effect of distance measures on accuracy for *Sph_5_2*, *Sph_6_2*, iris, wine, bupa and CMC datasets respectively. We have carried out 10 independent runs of each clustering algorithm, each run consists of 100 iterations. Each value mentioned in the tables is an average of 10 runs.

Table 1. Effect of distance measures on accuracy of cluster formed for *Sph_5_2* dataset

	<i>Euclidean</i>	<i>Seuclidean</i>	<i>Manhattan</i>	<i>Mahalanobis</i>	<i>Cosine</i>	<i>Correlation</i>
<i>SingleLinkage</i>	0.5960	0.5880	0.3560	0.5880	0.4640	0.4000
<i>AverageLinkage</i>	0.9400	0.9480	0.9440	0.9480	0.5080	0.4040
<i>K - Means</i>	0.9575	0.9650	0.8625	0.9590	0.5825	0.4000

For *Sph_5_2* (Table 1), it is found that the single linkage with Euclidean distance provides better accuracy over other distance measures. The similar results are found for both hierarchical algorithms with Standard Euclidean and Mahalanobis distance. The K-means clustering algorithm with Mahalanobis distance attains best accuracy. The single linkage and average linkage clustering

Table 2. Effect of distance measures on accuracy of cluster formed for *Sph_6_2* dataset

	<i>Euclidean</i>	<i>Seuclidean</i>	<i>Manhattan</i>	<i>Mahalanobis</i>	<i>Cosine</i>	<i>Correlation</i>
<i>SingleLinkage</i>	1.0000	1.0000	1.0000	1.0000	0.4833	0.3200
<i>AverageLinkage</i>	1.0000	1.0000	1.0000	1.0000	0.5766	0.3400
<i>K – Means</i>	0.7941	0.7392	0.7917	0.7995	0.5783	0.3250

Table 3. Effect of distance measures on accuracy of cluster formed for Iris dataset

	<i>Euclidean</i>	<i>Seuclidean</i>	<i>Manhattan</i>	<i>Mahalanobis</i>	<i>Cosine</i>	<i>Correlation</i>
<i>SingleLinkage</i>	0.6800	0.6600	0.6733	0.3533	0.6600	0.6600
<i>AverageLinkage</i>	0.9067	0.6867	0.9000	0.3467	0.6600	0.9467
<i>K – Means</i>	0.8442	0.8442	0.7667	0.7783	0.8033	0.8007

Table 4. Effect of distance measures on accuracy of cluster formed for Wine dataset

	<i>Euclidean</i>	<i>Seuclidean</i>	<i>Manhattan</i>	<i>Mahalanobis</i>	<i>Cosine</i>	<i>Correlation</i>
<i>SingleLinkage</i>	0.4270	0.3764	0.3989	0.3876	0.4101	0.3876
<i>AverageLinkage</i>	0.6124	0.3876	0.5449	0.3876	0.4475	0.4719
<i>K – Means</i>	0.6699	0.6699	0.6699	0.6088	0.6853	0.6853

Table 5. Effect of distance measures on accuracy of cluster formed for Bupa dataset

	<i>Euclidean</i>	<i>Seuclidean</i>	<i>Manhattan</i>	<i>Mahalanobis</i>	<i>Cosine</i>	<i>Correlation</i>
<i>SingleLinkage</i>	0.5768	0.5768	0.5710	0.5768	0.5768	0.5768
<i>AverageLinkage</i>	0.5565	0.5710	0.5623	0.5768	0.5623	0.5739
<i>K – Means</i>	0.5507	0.5510	0.5478	0.5341	0.5072	0.5362

Table 6. Effect of distance measures on accuracy of cluster formed for CMC dataset

	<i>Euclidean</i>	<i>Seuclidean</i>	<i>Manhattan</i>	<i>Mahalanobis</i>	<i>Cosine</i>	<i>Correlation</i>
<i>SingleLinkage</i>	0.4270	0.3938	0.4257	0.3938	0.4270	0.4277
<i>AverageLinkage</i>	0.4094	0.3938	0.4094	0.4257	0.4100	0.4257
<i>K – Means</i>	0.3976	0.3987	0.4010	0.3919	0.3795	0.3794

algorithm for *Sph_6_2* dataset, attain 100 percent accuracy with four distance measures (Euclidean, Standard Euclidean, Manhattan, and Mahalanobis). The best value is provided by K-Means with Mahalanobis distance.

For the iris dataset, (Table 3) the single linkage with Euclidean distance attains better accuracy than the other distance measures. However, the average linkage with correlation distance provides higher accuracy. The K-Means clustering algorithm provides better accuracy with Euclidean and standard Euclidean distance. For wine dataset, (Table 4) the single and average linkage with Euclidean distance find optimal accuracy. K-means with cosine distance provides better accuracy than the other distance measures.

The results obtained for the Bupa dataset (Table 5) show that the single linkage clustering produces similar accuracy for five distance measures except cosine. Average linkage using Mahalanobis distance attains best accuracy value. The K-Means clustering algorithm provides better accuracy with Euclidean and standard Euclidean distance. For CMC dataset results given in Table 6, show that the Single and average linkage clustering using correlation distance provide good accuracy. For K-Means, manhattan distance produces superior results.

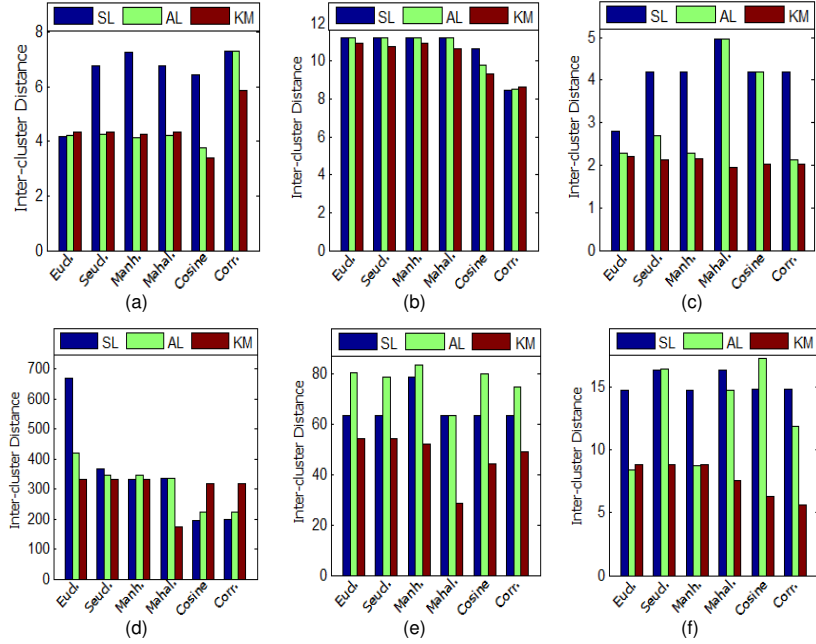


Fig. 1. Effect of distance measures on inter-cluster distance; (a) *Sph_5_2* (b) *Sph_6_2* (c) iris (d) wine (e) bupa (f) CMC.

Figure 1 shows the effect of distance measures on inter cluster distance for all above mentioned datasets respectively. For *Sph_5_2*, clustering using correlation distance provides well separated clusters. For *Sph_6_2*, the clustering algorithms with euclidean distance measure attains better cluster separation than the other distance measures. Single and average linkage with mahalanobis distance produce well separated clusters for iris dataset. For wine dataset, single and average linkage with euclidean distance provide better separation between clusters than the other distance measures. The K-Means with manhattan distance offers superior cluster separation over other distance measures. For bupa dataset, single and average linkage with manhattan distance produce best cluster separation. Single linkage with standardized euclidean gives better cluster separation than other distance measures for CMC dataset. For bupa and CMC datasets, K-Means with standardized euclidean gives well separated clusters.

5 Conclusions

The distance measures play a crucial role in clustering. In this paper, we have analyzed three well-known clustering algorithms by using six different widely used distance measures. We have used two artificial and four real-life datasets for experimentation. The experimental results reveal that different distance measures yield better values for different datasets when used for different algorithms. This paper can help in selecting suitable distance measure that can provide the optimum results.

References

1. Bandyopadhyay, S., Maulik, U.: Genetic clustering for automatic evolution of clusters and application to image classification. *Pattern Recogn.* 2, 1197–1208 (2002)
2. Chen, J., Zhao, Z., Ye, J., Liu, H.: Nonlinear adaptive distance metric learning for clustering. In: *ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pp.123–132. New York (2007)
3. Duda, R.O., Hart, P.E., Stork, D.G.: *Pattern Recognition*. John Wiley and Sons (2001)
4. Jain, A.K., Dubes, R.C.: *Algorithms for clustering data*. Prentice Hall, Englewood Cliffs, NJ (1988)
5. Kaufman, L., Rousseeuw, P.: *Finding groups in data: An introduction to cluster analysis*. John Wiley and Sons, New York (1990)
6. Mahalanobis, P.C.: On the generalized distance in statistics. In: *Naitional Institute of Sciences of India*, pp.49–55. Chicago (1936)
7. Prekopsak, Z., Lemire, D.: Time series classification by class-specific mahalanobis distance measures. *Advances in Data Analysis and Classification*, 6(3), 185–200 (2012)
8. Vimal, A., Valluri, S.R., Karlapalem, K.: An experiment with distance measures for clustering. Technical report, Center of Data Engineering, IIIT, Hyderabad (2008)
9. Xu, R., Wunsch, D.: Survey of clustering algorithms. *IEEE Trans. Neural Networks*. 16(3), 513–521 (2005)
10. UCI Repository of Machine Learning Databases, <http://www.ics.uci.edu/~mlearn/databases>