A COMPARATIVE STUDY AND PERFORMANCE EVALUATION OF SIMILARITY MEASURES FOR DATA CLUSTERING

[1*]DAUDA USMAN AND [2]ISMAIL BIN MOHAMAD

[1,2]Department of Mathematical Sciences, Faculty of Science
Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia
[1*]dauusman@gmail.com, [2]ismailm@utm.my
*Corresponding author

Abstract. Clustering is a useful technique that organizes a large quantity of unordered datasets into a small number of meaningful and coherent clusters. A wide variety of distance functions and similarity measures have been used for clustering, such as squared Euclidean distance, Manhattan distance and relative entropy. In this paper, we compare and analyze the effectiveness of these measures in clustering for high dimensional datasets. Our experiments utilize the basic *K*-means algorithm with application of PCA and we report results on simulated high dimensional datasets and two distance/similarity measures that have been most commonly used in clustering. The analyzed results indicate that Squared Euclidean distance is much better than the Manhattan distance method.

Keywords basic k-means, clustering technique, manhattan, outliers, similarity measures, squared euclidean.

1.0 INTRODUCTION

Clustering is a process of grouping a set of physical objects into classes of *similar* objects and is a most interesting concept of data mining in which it is defined as a collection of data objects that are similar to one another. Purpose of Clustering is to catch fundamental structures in data and classify them into meaningful group. Many of the clustering algorithms have been published every year and can be proposed for different research fields. They were developed by using various techniques and approaches. But according to the recent study *K*-means has been one of the top most data mining algorithms presently. For many of the practitioners k-means is the favorite algorithm in their related fields to use. Even though it is a top most algorithm, it has a few basic drawbacks when clusters are of differing sizes, densities and non-globular shape. Irrespective of the drawbacks its simplicity, understandability, and scalability is the main reasons that made the algorithm popular.

An algorithm with adequate performance and usability in most of application scenarios could be preferable to one with better performance in some cases but limited usage due to high complexity. While offering reasonable results, K-means is fast and easy to combine with other methods in larger systems. Hartigan [1] opined that cluster analysis is one tool that is used in the exploration of data in which the interactions among patterns are assessed by placing them into groups with unique and distinct characteristics. Later, [2] defined cluster analysis as a technique for creating groups of objects such that each cluster contains points that are similar and unique.

The objective is targeted at finding the best grouping for which the observations or objects found in within each cluster are the same. "More accurately, cluster analysis consists of a series of processes that partition a given data set $X = \{\vec{x}_1, \vec{x}_2\ , \dots, \vec{x}_n\} \subset \mathcal{R}^D$ into clusters such that the data points in a cluster are more similar to each other than points in different clusters" [3]. Thus the principal interest in the clustering process is the revelation of sensible groups or patterns, which allow for the discovery of similarities and dissimilarities so that useful conclusions can be reached.

Basically, there is an implicit assumption that the true intrinsic structure of data could be correctly described by the similarity formula defined and embedded in the clustering criterion function. Hence, effectiveness of clustering algorithms under this approach depends on the appropriateness of the similarity measure to the data at hand. The work in this paper is motivated by investigations from the above and similar research findings. It appears to us that the nature of similarity measure plays a very important role in the success or failure of a clustering method. Hence, our objective is to check the best method for measuring similarity between data objects in sparse and high-dimensional domain which is fast, capable of providing high quality clustering result and consistent performance.

2.0 MATERIALS AND METHODS

Before clustering, a similarity/distance measure must be determined. The measure reflects the degree of closeness or separation of the target objects and should correspond to the characteristics that are believed to distinguish the clusters embedded in the data. In many cases, these characteristics are dependent on the data or the problem context at hand, and there is no measure that is universally best for all kinds of clustering problems. Moreover, choosing an appropriate similarity measure is also crucial for cluster analysis, especially for a particular type of clustering algorithms. For example, the density-based clustering algorithms, such as DBScan rely heavily on the similarity computation.

Therefore, understanding the effectiveness of different measures is of great importance in helping to choose the best one. However, not every distance measure is a metric. Also to qualify as a metric, a measure d must satisfy the following four conditions: Let x and y be any two objects in a set and $d(x, y)$ be the distance between x and y.

1. The distance between any two points must be nonnegative, that is, $d(x, y) \geq 0$.

2. The distance between two objects must be zero if and only if the two objects are identical, that is, $d(x, y) = 0$ if and only if $x = y$.

3. Distance must be symmetric, that is, distance from x to y is the same as the distance from y to x, ie. $d(x, y) = d(y, x)$.

4. The measure must satisfy the triangle inequality, which is $d(x, z) \leq d(x, y) + d(y, z)$.

2.1 Similarity Measures

Similarity measures quantify how "similar" two patterns are. In most cases we have to ensure that all selected features contribute equally to a similarity measure and there are no features that dominate others. Similarity is fundamental to the definition of a cluster; a measure of the similarity between two patterns drawn from the same feature space is essential to most clustering procedures. It is most common to calculate the dissimilarity between two patterns using a distance measure defined on the feature space. Because of the variety of feature types and scales, the distance measures must be chosen carefully.

Distances and similarities play an important role in cluster analysis [4, 5]. In the literature of data clustering, similarity measures, similarity coefficients, dissimilarity measures, or distances are used to describe quantitatively the similarity or dissimilarity of two data points or two clusters.

In general, distance and similarity are reciprocal concepts. Often, similarity measures and similarity coefficients are used to describe quantitatively how similar two data points are or how similar two clusters are: the greater the similarity coefficient, the more similar are the two data points. Dissimilarity measure and distance are the other way around: the greater the dissimilarity measure or distance, the more dissimilar are the two data points or the two clusters.

Every clustering algorithm is based on the index of similarity or dissimilarity between data points [4]. If there is no measure of similarity or dissimilarity between pairs of data points, then no meaningful cluster analysis is possible. A distance metric is a real-valued function *d*, such that for any points *x, y* and *z:*

$d(x, y) \geq 0$, and $(x, y) = 0$ if and only if $x = y$
(1.1)
$d(x, y) = d(y, x)$
(1.2)
$d(x, z) \leq d(x, y) + d(y, z)$ (1.1)

First property, positive definiteness, assures that distance is always a nonnegative quantity, so the only way distance can be zero is for the points to be the same. The second property indicates the symmetry nature of distance. The third property is the triangle inequality, according to which introducing a third point can never shorten the distance between two points [6]. There are several measures of distance which satisfy the metric properties, some of which are:

2.2 Euclidean Distance

The Euclidean distance is the most common distance metric used in low dimensional data sets. It is also known as $L_2 norm.$ The Euclidean distance is the usual manner in which distance is measured in real world. In this sense, Manhattan distance tends to be more robust to noisy data.

$$d_{euclidean}(X,Y) = \sqrt{\sum_i (x_i - y_i)^2} \qquad (1.2)$$

where X and Y are m-dimensional vectors and denoted by $X = (x_1, x_2, x_3, \cdots, x_m)$ and $Y = (y_1, y_2, y_3, \cdots, y_m)$ represent the m attribute values of two records [6]. While Euclidean metric is useful in low dimensions, it doesn't work well in high dimensions. The drawback of Euclidean distance is that it ignores the similarity between attributes. Each attribute is treated as totally different from all of the attributes [7].

2.3 Manhattan Distance

This metric is also known as $L_1 norm\ or\ the\ rectilinear\ distance.$ This is also a common distance metric and gets its name from the rectangular grid patterns of streets in midtown Manhattan. Hence, another name for the distance metric is also *city block distance*. It is defined as the sum of distances travelled along each axis.

The Manhattan distance looks at the absolute differences between the coordinates. In some situations, this metric is more preferable to Euclidean distance, because the distance along each axis is not squared so a large difference in one dimension will not dominate the total distance [8].

$$d_{manhattan}(X,Y) = \sum_i^m |x_i - y_i| \qquad (1.3)$$

3.0 EXPERIMENTAL RESULTS AND DISCUSSION

It is very difficult to conduct a systematic study comparing the impact of similarity metrics on cluster quality, because objectively evaluating cluster quality is difficult in itself. In practice, manually assigned category labels are usually used as baseline criteria for evaluating clusters. As a result, the clusters, which are generated in an unsupervised way, are compared to the pre-defined category structure, which is normally created by human experts. This kind of evaluation assumes that the objective of clustering is to replicate human thinking, so a clustering solution is good if the clusters are consistent with the manually created categories.

However, in practice datasets often come without any manually created categories and this is the exact point where clustering can help. Therefore, measures like cluster coherence in terms of the within-cluster distances and the well-separateness between clusters in terms of between-cluster distances can were used for evaluation in this paper.

3.1 Results

This section compares the two distance functions, as discussed in section 2. The K-mean clustering algorithm was implemented using each of the distance functions: Squared Euclidian and City Block distance measures. A simulation experiment is conducted to compare the cluster formations and the running time required by the two approaches. We generated n random data from multivariate distribution $N_p(\mu, \Sigma)$, where Σ is positive definite. In order to make the advantage of the two approaches very clear, show its separation and compactness the paper consider two and three centroids. The analysis was carried out with pair (p, n) = (20, 500), (50, 500) and 500 replicates for each run. By utilizing M-file Matlab 7.6 (R2008a), the required time taken and the last five steps of the sum squares errors of the two approaches are presented in Table 1 and their respective cluster formations shown in Figure 1 to 8 respectively.

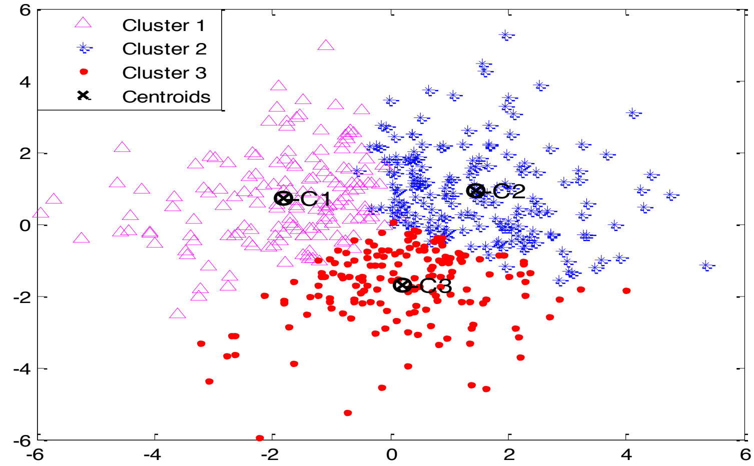

Figure 1: K-means clustering with SED

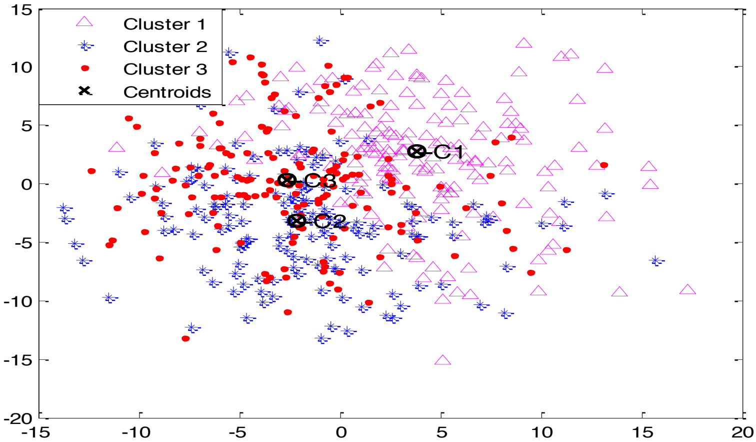

Figure 2: K-means clustering with MD

Figure 1 and 2 gives the results of the *K*-means clustering using Squared Euclidean distance (SED) and Manhattan distance (MD) with simulated dataset containing 500 sample size and 20 variables. Their error sums of squares are 14567.2, 35928.9 and the CPU time taken equal 9.63 and 10.45 respectively.

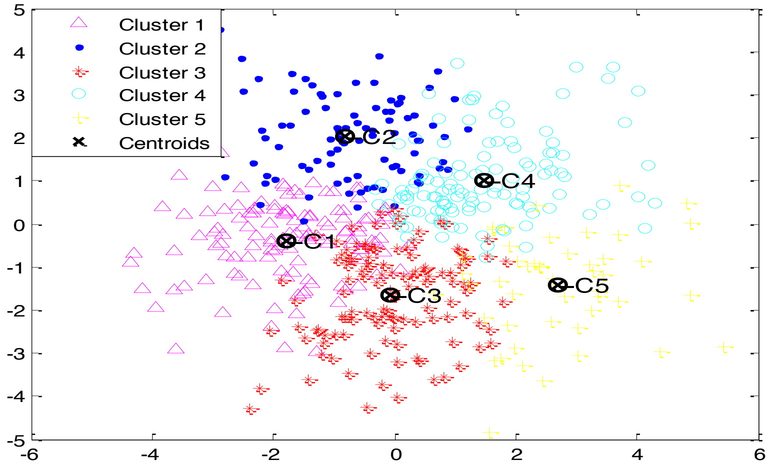

Figure 3: K-means clustering with SED

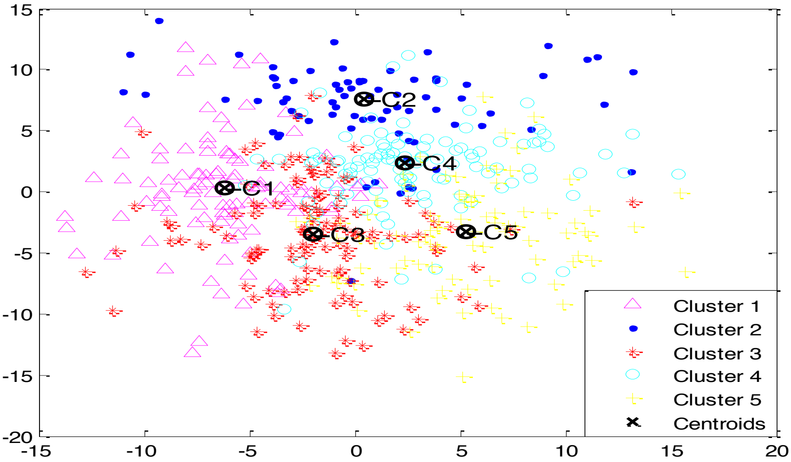

Figure 4: K-means clustering with MD

Figure 3 and 4 gives the results of the *K*-means clustering using Squared Euclidean distance (SED) and Manhattan distance (MD) with simulated dataset containing 500 sample size and 20 variables. Their error sums of squares are 13948.5, 34918.5 and the CPU time taken equal 6.74 and 7.16 respectively.

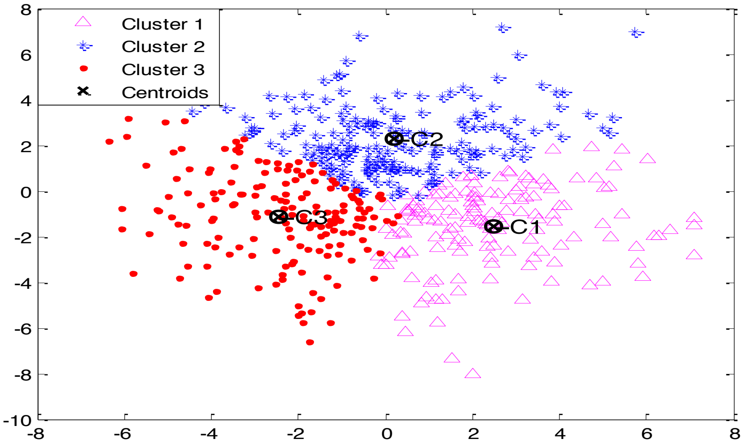

Figure 5: K-means clustering with SED

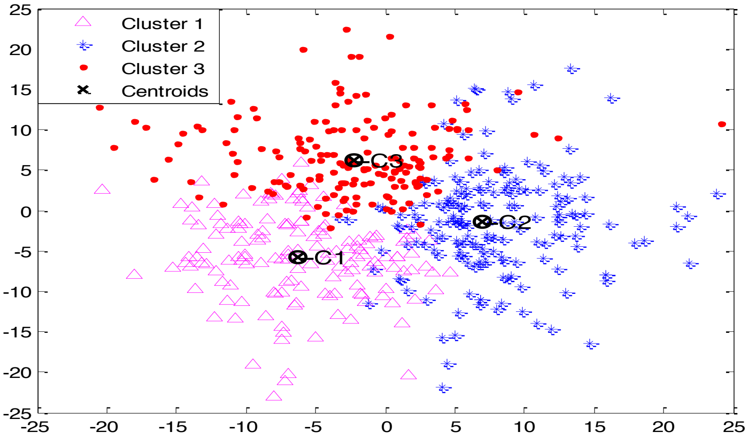

Figure 6: K-means clustering with MD

Figure 5 and 6 gives the results of the *K*-means clustering using Squared Euclidean distance (SED) and Manhattan distance (MD) with simulated dataset containing 500 sample size and 50 variables. Their error sums of squares are 28581.4, 61354.5 and the CPU time taken equal 07.86 and 09.34 respectively.

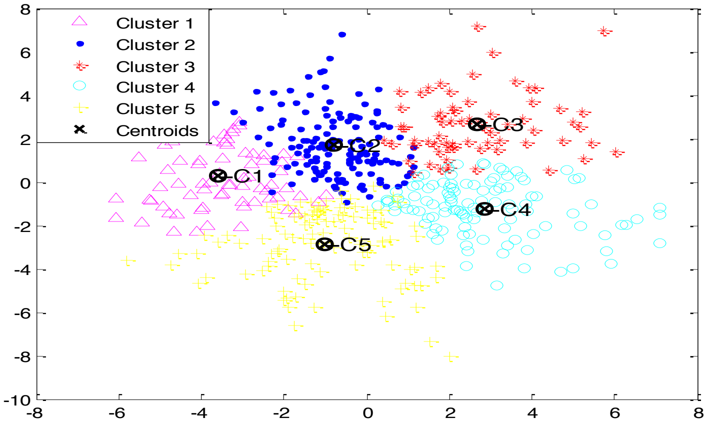

Figure 7: K-means clustering with SED

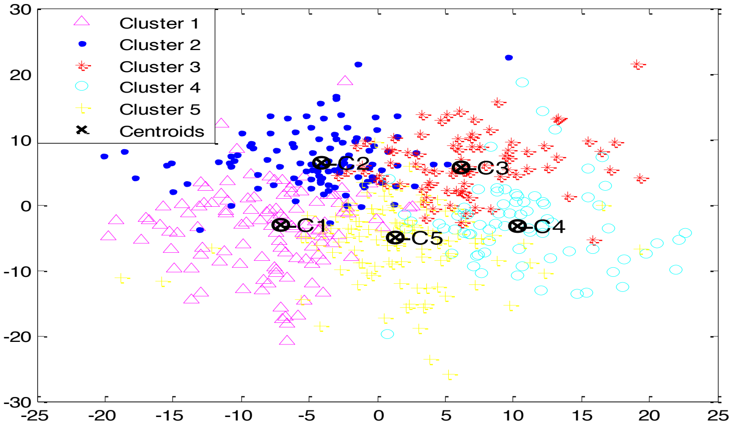

Figure 8: K-means clustering with MD

Figure 7 and 8 gives the results of the *K*-means clustering using Squared Euclidean distance (SED) and Manhattan distance (MD) with simulated dataset containing 500 sample size and 50 variables. Their error sums of squares are 27380.2, 60351.4 and the CPU time taken equal 08.11 and 09.89 respectively.

Table 1: Error Sum of Squares and Time Taken

Method	Error Sum of Squares (20, 500)	Time Taken (20, 500)	Error Sum of Squares (50, 500)	Time Taken (50, 500)
SED 3 Centers	14567.2	9.63	28581.4	07.86
MD 3 Centers	35928.9	10.45	61354.5	09.34
SED 5 Centers	13948.5	6.74	27380.2	08.11
MD 5 Centers	34918.5	7.16	60351.4	09.89

4.0 CONCLUSION

A distance measuring function is used to measure the similarity among objects, in such a way that more similar objects have lower dissimilarity value. Several distance measures can be employed for clustering tasks. Each measure has its own merit and demerits. The selection of different measures is a problem dependent. Hence, choosing an appropriate distance measure for *K*-mean clustering algorithm can greatly reduce the burden of the algorithm. The performance of the Squared Euclidean distance measure outperforms the Manhattan distance measure. The results reveal that the *K*-means algorithm with Squared Euclidean distance measure accurately maximizes the cluster accuracy and can play a critical role for both low and high dimensional dataset.

REFERENCES

[1] J. Hartigan, and M. Wang (1979). A K-means clustering algorithm. Appl. Stat., 28:100-108.

[2] Guojun, G., Chaoqun, M. and Jianhong, W. (2007). *Data Clustering Theory, Algorithms and Applications*. American Statistical Association and The Society for Industrial and Applied Mathematics.

[3] C. Moses, S. Guha, E. Tardos and B. S. David (1999). A Constant-Factor Approximation Algorithm for the *K*-median Problem. Proceedings of the 31st Annual ACM Symposium on Theory of Computing.

[4] A. Jain, and R. Dubes (1988). *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice–Hall.

[5] M. Anderberg, (1973). *Cluster analysis for applications*. New York: Academic Press. S. Salleh, S. Olariu and B. Sanugi. Single-row transformation of complete graphs. *Journal of Supercomputing*, 31, 265-279, 2005.

[6] D.T. Larose, (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*, New Jersey: John Wiley and Sons.

[7] L. Ertoz., M. Steinbach and V. Kumar (2003). Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data, *Proceedings of the Third SIAM International Conference on Data Mining, Volume 3*, 2003, San Francisco.

[8] M. J. A. Berry, and G. S. Linoff (1997). *Data Mining Techniques for Marketing, Sales and Customer Support*. John Wiley & Sons, Inc., New York.