CHAPTER 8

Clustering Algorithms using Different Distance Measures

8.1 Introduction

In this chapter[1], clustering algorithms such as k-means and k-medoids are described using various types of distance measures. With an illustration, the algorithms are derived and the results are compared. The data base which deals with medical science i.e. 400 records are related to the vision of the eyes. Cluster formation is discussed based on k-means with different distance measures and similarly for k-medoids. The results are discussed based on the comparison of two cluster algorithms. Cluster analysis is an approach that finds structure in data by identifying natural

[1] part of this work was published in CiiT International Journal Data Mining and Knowledge Engineering, Vol 5,no. 4, pp 140-143 (2013).

groupings in the data. Unfortunately natural groupings are not as well defined as they might hope. Indeed, it is usual to have more than one natural grouping for any collection of data. As we will see, there is no definitive cluster analysis technique, instead the term relates to a rather loose collection of algorithms that group similar objects into categories. Although some clustering algorithms have been present in standard statistical software packages for many years, they are rarely used for formal significance testing.

A cluster is simply a collection of cases that are more similar to each other than they are to cases in other clusters. This intentionally vague definition is common; for example, Sneath and Sokal (1973) noted that vagueness was inevitable given the multiplicity of different definitions while Kaufman and Rousseeuw (1990) referred to cluster analysis as the art of finding groups. Clustering divides a database into different groups. The goal of clustering is to find groups that are very different from each other, and whose members are very similar to each other. Unlike classification, you don't know what the cluster will be when you start, or by which attributes the data will be clustered.

Consequently, someone who is knowledgeable in the business must interpret the clusters. Often it is necessary to modify the clustering by excluding variables that have been employed to group instances, because upon examination the user identifies them as irrelevant or not meaningful. After you have found clusters that reasonably segment your database, these clusters may then be used to classify new data. Some of the common algorithms used to perform clustering include K-means and K-medoids.

Clustering and Segmentation are similar but not same methods. Segmentation refers to the general problem of identifying groups that have common characteristics. Clustering is a way to segment data into groups that are not previously defined, whereas classification is a way to segment data by assigning it to groups that are already defined.

Medical data exhibit certain features that make their classification stand out as a distinct field of research several medical classification tasks exist, among which medical diagnosis and prognosis are most common to present a review of medical data classification techniques, but rather to introduce a snapshot of data mining techniques used to aid medical decision making. There are four hundred diabetes patients taken from website www.healthdata.gov, on January- 2012 to March -2012. The variables are Body Mass Index, Waist Hip Ratio, Urine Glucoses, Blood Glucoses, Systolic, Diastolic and Years affected.Cluster analysis is the process of grouping a set objects into classes or clusters so that objects within a cluster have similar in comparison to one another, but there are dissimilar to objects in other clusters.

Two types of clustering are hierarchical and non-hierarchical method. These algorithms usually are either agglomerative or divisive. Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger cluster. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters. Partitional algorithms typically determine all clusters at once, but can also be used as divisive algorithms in the hierarchical clustering.

8.1.1 Categories of Clustering Algorithms

Clustering techniques can be divided into three main categories

1. Partition – based clustering, sometimes referred to as objective function-based clustering
2. Hierarchical clustering
3. Model- based (a mixture of probabilities) clustering

The clustering principles for each of these categories are very different which implies very different style of processing and resulting formats of the results. In Partition-based clustering rely on a certain objective function whose minimization is supposed to lead the discovery of the structure existing in the set. While the algorithmic setup is quite appealing and convincing, one is never sure what type of structure to expect and hence what should be most suitable form of the objective function. Typically, in this category of the methods that predefine the number of clusters and proceed with the optimization of the objective function.

There are some variants in which also allow for successive splits of the clusters, a process that leads to a dynamically adjusted number of clusters. The essence of hierarchical clustering lies in the successive development of clusters either with successive splits or with individual points treated as initial clusters, which is keep merging . The essential feature of hierarchical clustering concerns a suitable choice of a distance function and a means to express the distance between data and patterns. These features, in essence, give rise to a spectrum of various clustering methods such as single linkage, complete linkage etc. In model-based clustering, as the name itself stipulates a certain probabilistic model of the data and then estimate its parameters. It is also called as mixture density model where it assumes that the data are a result of mixture of c sources of data. Each of these sources is treated as a potential cluster.

8.1.2 Cluster validity

Normally, the user would guess the range of the number of clusters and then use the cluster validity measures to assess which particular number of clusters best reveals the true structure in the data. Cluster validity is a suite methodologies and algorithms that offer some mechanisms to validate clustering results. There are many measures called cluster validity indices, whose values relate to the number of clusters generated and thus are used to judge the clusters detected in the data and to assess the quality of the structure revealed in this manner.

Compactness:

This property expresses how close the elements in the clusters are. For instance, consider a variance of the elements; the lower the value of the variance, the higher the compactness of the cluster. In compact clusters, low values of compactness are desirable.

Separability:

For this property, evaluate how distinct the clusters are. An intuitive way of expressing separability is to compare inter cluster distances. Since for high compactness and high separability, a structure should be characterized by small values of intra cluster distances and large values of inter cluster distances.

8.2 Computation Algorithms

Cluster analysis is the process of grouping a set of objects into classes or clusters so that objects within a cluster have similar comparison to one another, but there are dissimilar objects in other clusters. In this section, two types of clustering algorithms are discussed such as k-means and k-medoids.

The first split in Grabmeier and Rudolph's (2002) taxonomy separated partitioning methods from hierarchical methods. Lingras and Huang (2005) suggest that partitioning methods belong to a class of cluster based methods, unlike the object based methods typified by hierarchical methods. The two classes relate to the method by which cluster characteristics are determined. In the object based methods the assignment of cases to clusters defines the clusters' characteristics while the cluster based assign weight vectors to the clusters.

K - Means Clustering

It was developed in 1967 by MacQueen and then modified 1975 by Hartigan and Wang. The k-means is one of the simplest unsupervised learning algorithms that solves the well known clustering problem. The k-means algorithm takes the input parameter, 'k', and the partitions, a set of 'n' objects into k clusters so that the resulting intracluster similarity is high but the intercluster similarity is low. Cluster similarity is measured in relation to the mean value of the objects in a cluster, which can be viewed as the cluster's centroid or center of gravity. The k-means algorithm which is one of the most frequently used is a remarkably simple and efficient algorithm. There are two steps that are repeated until a solution is found.

The algorithm partitions a set of N vector X = x_j j=1 to N into C classes v_i, i =1 to c and finds a cluster centre for each class c_i denotes the centroid of cluster v_i such that an objective function of dissimilarity, for example a distance measure, is minimized. The objective function that should be minimized, when the Euclidean distance is selected as a dissimilarity measure, can be described as follows

$$P = \Sigma_{i=1}^{c}(\Sigma_{k,xk}\|\|xk - ci\|)^2 \tag{8.1}$$

Where$\|xk - ci\|\|^2$ the objective is function within group i and $\|xk - ci\|^2$ is a chosen distance measure between a data point xk and the cluster centre ci.

The Partitioned groups are typically defined by a C x N binary membership matrix $U = (u_{ij})$, where the element u_{ij} is 1 if the j^{th} data point x_j belongs to group i, and o otherwise.

$U_{ij} = 1$ if $\|xj - ci\|^2 \leq \|xj - ck\|^2$ $\forall k \neq i$, 0 otherwise

$$C_i = \frac{\Sigma_{j=1,xj\epsilon ci}^{N} xj}{Ri} \tag{8.2}$$

Where R_i is number of data point in class v_i, since k-means method aims to minimize the sum of squared distances from all points to their cluster centers, this should result in compact clusters. We use the intracluster distance measure, which is simply the median distance between a point and its cluster centre. The equation is given as:

$$Intro = median\{\Sigma_{i=1}^{c}\Sigma_{k,xk}\|\|\|xk - ci\|^2\} \tag{8.3}$$

Therefore, the clustering which gives a minimum value for the validity measure will tell us what the ideal value of k is in the k-means. Then the number of cluster is known before estimating the membership matrix. The proposed k-means clustering algorithm is described as follows:

Computation algorithm for given k, k-means

Step 1: Partition objects into 'k' nonempty subsets

Step 2: Compute the seed points as the centroids of the clusters of the current partition (the centroid is the center, i.e., mean point, of the cluster)

Step 3: Assign each object to the cluster with the nearest seed point

Step 4: Go back to step 2, stop when no more new assignment.

Finally, this algorithm aims minimizing an objective function, in this case a mean squared error function is calculated as:

$$E = \Sigma_{i=1}^{k} \Sigma_{p \epsilon C_i} |p - m_i|^2 \tag{8.4}$$

Where E is the sum of the square error for all objects in the data set; P is the point in space representing a given object; and m_i is the mean of cluster C_i (both P and m_i are multidimensional).

K - Medoid Clustering

Partitioning Around Medoids (PAM) was introduces by Kaufman and Rousseeuw (1990). To find k clusters, PAMs approach is to determine a representative object for each cluster. This representative object called a medoid, is meant to be the most centrally located object within the cluster. Once the medoids have been selected each non-selected object is grouped with the medoid to which it is the most similar.

The objective of K- Medoid clustering is to find a non-overlapping set of clusters such that each cluster has a most representative point, i.e., a point that is the most centrally located with respect to some distance measure. These representative points are called medoids.The input of k-medoids algorithm is a dataset consisting of n vectors and it outputs k-clusters which together form a mutually exclusive and exhaustive partitioning of the data set.

The k-medoids algorithm expects the user to specify k, namely the number of clusters to be formed, this is the main reason for classifying k-medoids algorithm as a supervised clustering algorithm. This algorithm falls under the category of protoype-based algorithms, typically identified objects identifying ultimate clusters.

To compute the non-overlappind set of clusters k-medioids defined in the following steps:

Step 1: Select K initial points.

Step 2: Consider the effect of replacing one of the selected objects (mediods) with one of the non-selected objects.

Step 3: Select the configuration with the lowest cost.

Step 4: Otherwise, associate each non-selected point with its closest selected point (medoid) and stop.

Fuzzy clustering

In fuzzy clustering, objects are assigned to a particular cluster; they posses a membership function indicating the strength of membership in all or some of the clusters. The existing clustering techniques defined the 'strength of membership' has been either zero or one, with an object being either in or not in a cluster. Fuzzy clustering has two main advantages over crisp methods. Firstly, memberships can be combined with other information. In particular, in the special case where memberships are probabilities, results can be combined from different sources using

Bayes' theorem. Secondly, the memberships for any given object indicate whether there is a 'second best' cluster that is almost as good as the 'best' cluster, a phenomenon which is often hidden when using other clustering techniques.

$$J_m = \Sigma_{i=1}^{N} \Sigma_{j=1}^{C} u_{ij}^{m} ||x_j - c_j|| \tag{8.5}$$

- C is the total number of clusters,
- N is the total number of data, m is any real number greater than 1,
- w_{ij} is the degree of membership of x_j to the j^{th} cluster,x_j is the i^{th} of the d-dimensinonal measured data.
- c_j is the d-dimensional center of the cluster, and $|*|$ is any norm expressing the similarity between any measured data and the center.

Computation algorithm are as follows:

Step 1: Select on initial fuzzy fuzzy pseudo-partition, i.e., assign values to all the wij. (Like usual this can be done randomly or in a variety of ways.)

Step 2: Update the cluster centroids

Step 3: Update the w_{ij}.

Step 4: If the change in the error is below a specified threshold or the absolute change in any wij is below a given threshold, then stop go to step 2.

In fuzzy clustering analysis, the number of subsets is assumed known, and the membership function of each object in each cluster is estimated using an iterative method, usually a standard optimization technique based on heuristic objective function. In general, membership functions do not obey the rules of probability theory, although, once found, memberships can be scaled to lie between zero or one, and can then be interpreted as probabilities. The concept of a membership function derives from fuzzy logic, an extension of Boolean logic in which the concepts of true and false are replaced by that of partial truth. Boolean logic can be represented by set theory, and in an analogues manner fuzzy logic is represented by fuzzy set theory, Such techniques were originally developed for the description of natural language (Zadeh, 1965).

8.3 Distance Measures

The distances are normally used to measure the similarity or dissimilarity between two data objects. Some of the distance measures are Euclidean, Squared Euclidean and Manhattan.

Though there are various distance measures available in the literature. In this research work, the existing distance measures such as Euclidean, Manhattan distance and Squared Euclidean distance are applied on medical data and a comparative analysis is made.

In the playing card example clusters were formed by placing similar cards into the same cluster. All clustering algorithms begin by measuring the similarity between the cases to be clustered. Cases that are similar will be placed into the same cluster. It is also possible to view similarity by its inverse, the distance between cases, with distance declining as similarity increases. This leads to a general conclusion that objects in the same cluster will be closer to each other than they are to objects in other cluster. It also means that there must some means of measuring distance.

Euclidean Distance

The Euclidean distance between two data points involves computing the square root of the sum of the squares of the differences between corresponding values.

It represents the Euclidean distance when q=2,

$$d = (i, j) = \sqrt{\Sigma_{k=1}^{n}[X_{ik} - X_{jk}]^2} \tag{8.6}$$

Where $i = (X_{i1}, X_{i2}, \ldots, X_{in}) and j = (X_{j1}, X_{j2}, \ldots, X_{jn})$ are two n dimensional data objects.

Squared Euclidean Distance

The Squared Euclidean distance metric uses the same equation as the Euclidean distance metric, but does not take the squared root.

$$d(i,j) = \Sigma_{k=1}^{n}(X_{ik} - X_{jk})^2 \tag{8.7}$$

Squared Euclidean distance the sum of the squared differences between scores for two cases on all variables, i.e. the squared length of the hypotenuse. This measure magnifies distance between cases that are further apart.

Manhattan Distance

It is also known as City block distance, and absolute value distance or L_1 distance. Manhattan distances a distance that follows a route along the non-hypotenuse sides of a triangle. The name refers to the grid-like layout of most American cities which makes it impossible to go directly between two points. This metric is less affected by outliers than the Euclidean and squared Euclidean metries:

$$d(i,j) = \Sigma_{k=1}^{n}|X_{ik} - X_{jk}| \tag{8.8}$$

Mahalanobis Distance

A generalized version of a Euclidean distance which weights variables using the sample variance-covariance matrix. Because the covariance matrix is used this also means that correlations between variables are taken into account.

$$Distance(a,b) = [(a_i,b_i)^t S^{-t}(a_i - b_i)] \tag{8.9}$$

where S^{-1} is the inverse covariance matrix.

Minkowski Distances

Minkowski distance is a generalization of both Euclidean distance and Manhattan distance. It is defined as:

$$d(ij) = \mathit{\Sigma}(|x_{ij} - x_{jk}|)^{1/q} \tag{8.10}$$

where q is a positive integer. The Minkowski distance is a metric as a result of the Minkowski inequality.

Chebyshev Distance

Chebyshev distance or (Tchebyshev distance), maximum metric, or L_∞ metric is a metric defined on a vector space where the distance between two vectors is the greatest of their differences along any coordinate dimension. The Chebyshev distance between two vectors is

$$d_{ij} = max|x_{ik} - x_{jk}| \tag{8.11}$$

The above distance satisfies the following the following mathematic requirements of a distance function.

Cosine Distance

This is a type of Pearson measure which considers the relative difference $(e.g. A \times B/|A|.|B|)$, assuming that the scale is uniform (that the distance from zero is relative). In some case, this gives better results, particularly where the data is not normally distributed. It is a measure of similarity between two vectors of n dimensions by finding the cosine of the angle between them, often used to compare documents in text mining. In addition, it is used to measure cohesion within cluster in the field of data mining. Given two vectors of attributes A and B, the cosine similarity θ is obtained using a dot product and magnitude as shown in equation

$$Similarity = cos(\theta) = \frac{AB}{|A||B|} \tag{8.12}$$

Bray Curties distance

Bray Curtis distance is also called Sorensen distance. It is a normalization method that is commonly used in botany, ecology and environmental science field. It views the space as grid similar to the city block distance. The Bray Curtis distance has a property that if all coordinates are positive, the value of the distance is between zero and one. Zero Bray Curtis distance represent exact similar coordinate. If both objects are in the zero coordinates, the Bray Curtis distance is undefined. The normalization is done using absolute difference divided by the summation as shown in equation

$$d_{ij} = \frac{\Sigma_{k=1}^{n}|x_{ik} - x_{jk}|}{\Sigma_{k=1}^{n}(x_{ik} + x_{jk})} \tag{8.13}$$

Canberra Distance

Canberra distance examines the sum of series of a fraction difference between coordinates of a pair of objects. Each term of fraction difference has value between 0 and 1. If one of the coordinate is zero, the term becomes unity regardless of the other value. If bot coordinates are zeros, the distance need to be defined as (0/0=0). Otherwise, the distance will take infinite value. This distance is very sensitive to a small change when both coordinates are near to zero. The Canberra distance is found using equation

$$d_{ij} = \Sigma_{k=1}^{n} \frac{|x_{ik} - x_{jk}|}{|x_{ik}| + |x_{jk}|} \tag{8.14}$$

8.4 Computational Results

A perusal of table 8.1 reveals that the values for the formation cluster centers. The computed values in the table are the means for each variable within each formation cluster. The three distance measures such as Euclidean, Manhattan and squared Euclidean distances are applied. The comparison of these distance measures are determined by within clusters sum of squares by clusters using histograms. The results are described as follows.

TABLE 8.1: Formation of Cluster K-Means

	Cluster		
	1	2	3
BMI	23.88	25.31	22.68
WHR	.84	.84	.82
UG	2	2	4
BG	269	165	415
Systolic	150	141	137
Diastolic	94	90	88
Yrs affected	6	4	4

TABLE 8.2: Comparison of Distance measures using K-means

	Euclidean	Manhattan	Squared
Cluster1	102.1967	233.37	10444.16
Cluster2	186.8526	645.28	34913.88
Cluster3	166.35	507.37	2772.32

It may be noted from table 8.2, that the values for comparison of distance measures using k-means. Then, Euclidean is the lowest values followed by Manhattan and squared distance respectively. In figure 8.1 indicated the same phenomena graphically.

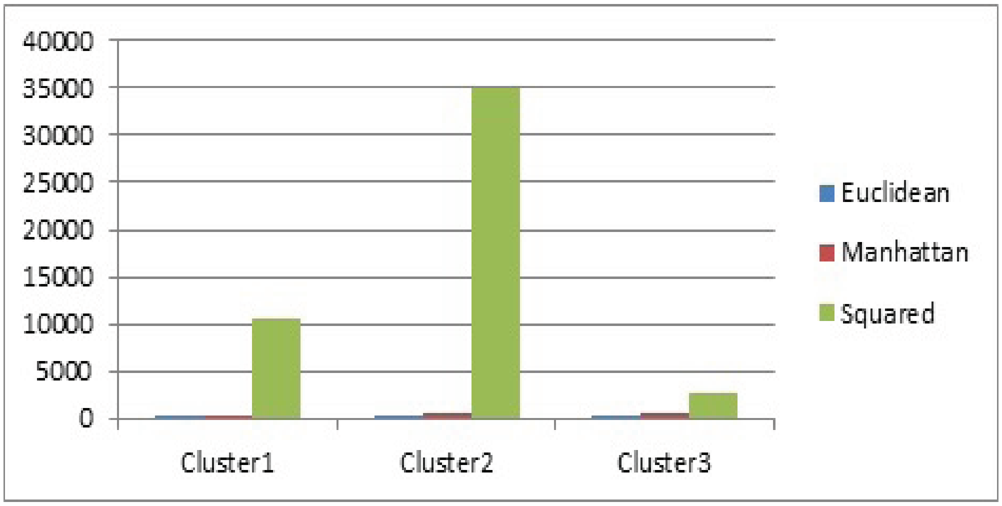

FIGURE 8.1: Distances Measures using K-Means

In table 8.3 reveals that the values which are used for the formation of cluster centers. The calculated values in the table are the means for each variable within each configuration cluster.

Euclidean, Manhattan and Squared Euclidean distance of three distances are applied. The comparison of this distance is determined by

Table 8.3: Formation of Cluster K-Medoids

Cluster			
	1	2	3
BMI	28.8	26.3	22.2
WHR	0.86	0.90	0.98
UG	1	0	3
BG	201	144	301
Systolic	120	166	150
Diastolic	80	95	90
Yrs affected	3	4	5

within clusters sum of squares by clusters using histograms. The results are defined as follows.

Table 8.4: Comparison of Distance Measures using k-medoids

	Euclidean	Manhattan	Squared
Cluster1	31.277	85.276	1349.871
Cluster2	27.276	74.699	1958.747
Cluster3	67.462	76.462	7177.909

Table 8.4 gives that the comparison of distance measures using k-medoids indicates the higher distance of the area have squared and followed Manhattan.

In the view of understanding, Figure 8.2 describes the distance measures using k-means graphically.

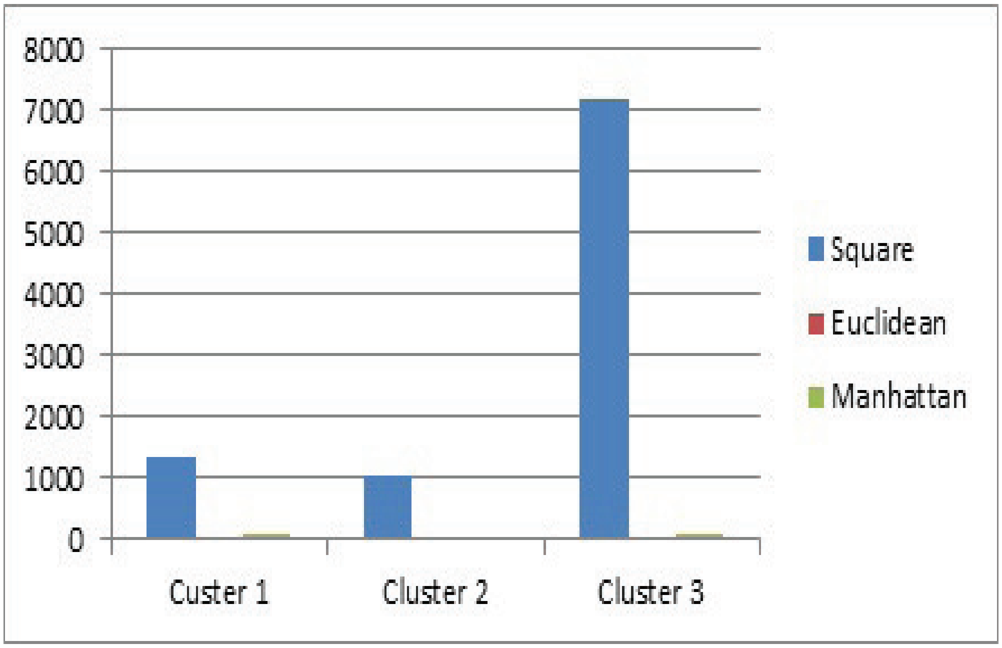

FIGURE 8.2: Distances Measures using k-means

8.5 Conclusion

In this chapter, the database consisting of 400 patients in this record. Out of 400 patients 120 patients have good vision, 240 have short sighted vision and 40 patients have blurred vision. Clusters are formed based on patients vision ailments using k-means and k-medoids clustering algorithms. Clusters algorithms are formed using different distance measures like Euclidean, Manhatten and Squared. The Euclidean distance measure is used in the k-means clustering outperforms with other two distance measures. Similarly, Euclidean distance measure is used in the k-medoids outperforms with other two distance measures. The comparative results show that the k-medoids clustering with Euclidean distance measure forms more densed cluster than the k-means clustering with Euclidean distance measure.