International Journal of Scientific Research in Science, Engineering and Technology (www.ijsrset.com)
© 2020 IJSRSET | Volume 7 | Issue 3 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
DOI : https://doi.org/10.32628/IJSRSET2073118

# Euclidean, Manhattan and Minkowski Distance Methods For Clustering Algorithms

**Aye Aye Thant, Soe Moe Aye**
Information Technology Supporting and Maintenance Department, Computer University (Mandalay), Mandalay, Myanmar

## ABSTRACT

The process of grouping a set of physical objects into classes of similar objects is called clustering. Clustering is a process of grouping the data into classes or cluster so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Dissimilarities are assessed based on the attribute values describing the objects. This system studies how to compute dissimilarities between objects represented by interval scaled variables. This system is intended to implement the dissimilarity matrix for interval-scaled variables using Euclidean, Manhattan, and Minkowski distance methods. This stores a collection of proximities that are available for all pairs of n objects.

**Keywords :** Clustering, dissimilarities, interval-scaled variables, Euclidean, Manhattan, Minkowski

## I. INTRODUCTION

Rapid advances in data collection and storage technology have enabled organization to accumulated vast amounts of data. However, extracting useful information has proven extremely challenging. Often, traditional data analysis tools and techniques cannot be used because of the massive size of a data set. Sometimes, the non-traditional nature of the means that traditional approaches cannot be applied even if the data set is relatively small. In other situations, the questions that need to be answered cannot be addressed using existing data analysis techniques, and thus, new methods need to be developed.

Data mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data. It has also opened up exciting opportunities for exploring and analysing new types of data and for analysing old types of data in new ways.

Traditional data analysis methods often deal with data sets containing attributes of the same type, either continuous or categorical. As the role of data mining in business, science, medicine, and other fields has grown, so has the need for techniques that can handle heterogeneous attributes.

Data sets differ in a number of ways. For example, the attributes used to describe data objects can be of different types- quantitative or qualitative- and data sets may have special characteristics; e.g., some data sets contain time series or objects with explicit relationships to one another.

Data is often far from perfect. While most data mining technique can tolerate some level of imperfection in the data, a focus on understanding and improving data quality typically improves the quality of the resulting analysis. Data quality issues that often need to be addressed include the presence

of noise and outliers; missing, inconsistent, or duplicate data; and data that is biased or, in some other way, unrepresentative of phenomenon or population that the data is supposed to describe. [3]
Often, the raw data must be processed in order to make it suitable for analysis. While one objective may be to improve data quality, other goals focus on modifying the data so that it better fits a specified data mining techniques or tool. For example, a continuous attributes, e.g., short, medium, or long, in order to apply a particular technique. As another example, the number of attributes in a data set is often reduced because many techniques are often reduced because many techniques are more effective when the data has a relatively small number of attributes.

One approach to data analysis is to find relationships among the data objects and then perform the remaining analysis using these relationships rather than the data objects themselves. For instance, can compute the similarity or distance between pairs of objects and then perform the analysis clustering, classification, or anomaly detection-based on these similarities or distances. There are many such similarity or distance measures, and the proper choice depends on the type of data and the particular application. [4]

## II. DISSIMILARITY MATRIX

Clustering is the process of grouping the data into classes or clusters so those objects within a cluster have high clusters. Dissimilarities are assessed based on the attribute values describing the objects. Often, distance measures are used. Clustering has its roots in many areas, including data mining, statistics, biology, and machine learning. [2]

### A. Similarity and Dissimilarity Measures

Similarity and dissimilarity are important because they are used by a number of data mining techniques, such as clustering, nearest, neighbour classification, and anomaly detection. In many cases, the initial data set is not needed once these similarities or dissimilarities have been computed. Such approaches can be viewed as transforming the data to a similarity (dissimilarity) space and then performing the analysis. The term proximity is used to refer to either similarity or dissimilarity. Since the proximity between two object is a function of the proximity between the corresponding attributes of the two objects, it first describe how to measure consider proximity measures for objects with multiple attributes.

This includes measures such as correlation and Euclidean distance, which are useful for such as time series or two-dimensional points, as well as the Jaccard and cosine similarity measures, which are useful for sparse data like documents.

Informally, the similarity between two objects is a numerical measure of the degree to which the two objects are alike. Consequently, similarities are higher for pairs of objects that are more alike. Similarities are usually non-negative and are often between 0 (no similarity) and 1 (complete similarity).

The dissimilarity between two objects is a numerical measure of the degree to which the two objects are different. Dissimilarities are lower for more similar pairs of objects. Frequently, the term distance is used as a synonym for dissimilarity, although, as it shall see, distance is often used to refer to a special class of dissimilarities. Dissimilarities sometimes fall in the interval [0, 1].

### B. Type of Data in Clustering

Suppose that a knowledge set to be clustered contains n objects, which can represent persons, houses, documents, countries, and so on. Main memory-based clustering algorithms typically operate either of the subsequent two data structures.

Data matrix (or object-by-variable structure): This represents n objects, like persons, with variables (also called measurements or attributes), like age, height, weight, gender, race, and so on. The structure is within the sort of a relational table, or n-by-p matrix (n objects * p variables):

**Dissimilarity matrix (or object-by-object structure): This stores a set of proximities that are available for all pairs of n objects. It is often represented by an n-by-n table.**

$$\begin{bmatrix} X_{11} & X_{1f} & X_{1p} \\ : & : & : \\ X_{i1} & X_{if} & X_{ip} \\ : & : & : \\ X_{np} & X_{n1} & X_{nf} \end{bmatrix}$$

$$\begin{bmatrix} 0 & & \\ d(2,1) & 0 & \\ d(3,1) & d(3,2) & 0 \\ : & : & : \\ d(n,1) & d(n,2) & \ldots\ldots\ldots\ldots 0 \end{bmatrix}$$

Where d (i, j) is that the measured difference or dissimilarity between objects i and j. In general, d (i, j) may be a nonnegative number that's on the brink of 0 when objects i and j are highly similar or "near" one another , and becomes larger the more they differ. Since d (i, j) = d (j, i) = 0.

The data matrix is often called a two-mode matrix, whereas the dissimilarity matrix is called a one-mode matrix, since the rows and columns of the former represent different entities, while those of the latter represent the same entity. Many clustering algorithms operate on a dissimilarity matrix. If the info are presented within the sort of a knowledge matrix, it can first be transformed into a dissimilarity matrix before applying such clustering algorithms.

Object dissimilarity can be computed for objects described by interval-scaled variables; by binary variables; by nominal, ordinal, and ratio-scaled variables; or combinations of these variable types. The dissimilarity data can later be used to compute clusters of objects. [1]

### C. Interval-Scaled Variables

Interval-scaled variables are continuous measurements of a roughly linear scaled. The measurement unit used can affect the clustering analysis. In certain applications, the dissimilarity (or similarity) between the objects described by interval-scaled variables is usually computed supported the space between each pair of objects. In general, expressing a variable in smaller units will lead to a larger range for that variable, and thus a larger effect on the resulting clustering structure. To help avoid dependency on the choice of measurement units, the data should be standardized. Standardizing measurements attempts to offer all variables an equal weight. This is particularly useful when given no prior knowledge of the info. However, in some applications, users may intentionally want to give more weight to a certain set of variables than to others.

### D. Distance Measures

The joining or tree clustering method uses the dissimilarities (similarities) or distances between objects when forming the clusters. Similarities are a group of rules that function criteria for grouping or separating items. In the previous example the rule for grouping a number of dinners was whether they shared the same table or not. [5]

Those distances (similarities) can be based on a single dimension or multiple dimensions, with each dimension representing a rule or condition for

grouping objects, For example, if we were to cluster fast foods, we could take into account the number of calories they contain , their price, subjective rating of taste, etc.

The most straightforward way of computing distances between objects during a multi-dimensional space is to compute Euclidean distances. If we had a two-or three-dimensional space this measure is that the actual geometric distance between objects within the space (i.e., as if measured with a ruler). However, the joining algorithm doesn't "care" whether the distances that are "fed" thereto are actual real distances, or another derived measure of distance that's more meaningful to the researcher; and it is up to the researcher the right method for specific application.

### E. Euclidean Distance

This is probably the foremost commonly chosen sort of distance. It simply is that the geometric distance within the multidimensional space. Note that Euclidean (and squared Euclidean) distances are usually computed from data , and not from standardized data. This method has certain advantages (e.g., the space between any two objects isn't suffering from the addition of latest objects to the analysis, which can be outliers). However, the distances are often greatly suffering from differences in scale among the size from which the distances are computed. For example, if one of the dimensions denotes a measured length in centimeters, and you then convert it to millimeters (by multiplying the values by 10), the resulting Euclidean or squared Euclidean distances (computed from multiple dimensions) can be greatly affected (i.e., biased by those dimensions which have a larger scale), and consequently, the results of cluster analyzes may be very different. Generally, it's good practice to rework the size in order that they have similar scales. [1]

After standardized, or without standardization in certain applications, the dissimilarity (or similarity) between the objects described by interval-scaled variables is usually computed supported the space between each pair of objects.

The most popular distance measure is Euclidean distance, which is defined as

$d(i, j)=\sqrt{|xi1- xj1|2 + | xi2- xj2|2+….+| xip- xjp|2}$

where i = ( xi1, xi2, ……xip) and j = (xj1, xj2,….., xjp) are two p-dimensional data objects.

### 2.3. Manhattan Distance

The Manhattan distance function computes the distance that would be veled to get one data point to the other if a grid-like path is followed. The Manhattan distance between two items is the sum of the differences of the corresponding components. It is also known as City Block distance. It represents distance between points during a city road grid. It examines absolutely the differences between coordinates of a pair of objects.

The formula for this distance between a point X= (Xi1, Xi2, etc.) and another point X= (Xj1, Xj2.etc) is:

Another well-known metric is Manhattan (or city block) distance, defined as

$d(i, j)= |xi1- xj1| + | xi2- xj2| +….+| xip- xjp|$

where Xi and Xj are the values of the ith variable and jth variable at points X respectively.

Both the Euclidean distance and Manhattan distance satisfy the subsequent mathematic requirements of a distance function:

1. $d(i, j) >= 0$: Distance is a nonnegative number.
2. $d(i, i) = 0$: the space of an object to itself is 0.
3. $d(i j) = d(j, i)$: Distance is a symmetric function.
4. d (i, j)

### 2.4. Minkowski Distance

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

$d(i, j)=( |xi1- xj1|q + | xi2- xj2|q +…+| xip- xjp|q)1/q$

Where q is a positive integer. It represents the Manhattan distance when q = 1, and Euclidean

distance when q = 2. When q = 3, 4, 5 and ∞, the values of matrix are less. [1]

## Algorithm

Algorithm : Dissimilarity Matrix
Input : Selected record set from sample database D.
Output : Dissimilarity matrix
Method:

```
1. matrix Values = two dimensional array of n objects * p variables;
2. For (i = 0; i <D.columns.count; i ++)
3. For (j = 0; j <D.columns.count; j ++)
4. If (i = j)
5.  matrixValues [i] [i] = 0;
6. else if (j> i)
7. {
8. tmp = calcDistance (i, j, "Distance Method");
9. matrixValues [i] [i] = tmp;
10.  matrixValues [j] [i] = tmp;
11. }
```

Function: calcDistance (index1, index2, method)

```
1. distance = 0.0;
2. q = user defined q value;
3. For (i = 0; i <D.columns.count; i+ +)
4. {
5. if (method "Euclidean")
6. {
7. distance + = Math.Pow ((D.row [index1] .ItemArray [i] .value, D.row [index2 ] .ItemArray [i) .value), 2);
8. }
9. else if (method = "Manhattan")
10. {
11. distance + D.row [index1] .ItemArray [i] .value, D.row [index2] .ItemArray [i] .value;
12. }
13. else if (method = "Minkowski")
14. {
15. distance + = Math.Pow ((D.row [index1] .ItemArray [i] .value, D.row [index2] .ItemArray [i] .value), q);
16. }
17. }
18. if (method ="Euclidean")
19. distance Math.sqrt (distance);
20. else if (method = "Minkowski")
21. distance = Math.Pow (distance, 1/q);
22. return distance
```

## III. System Design

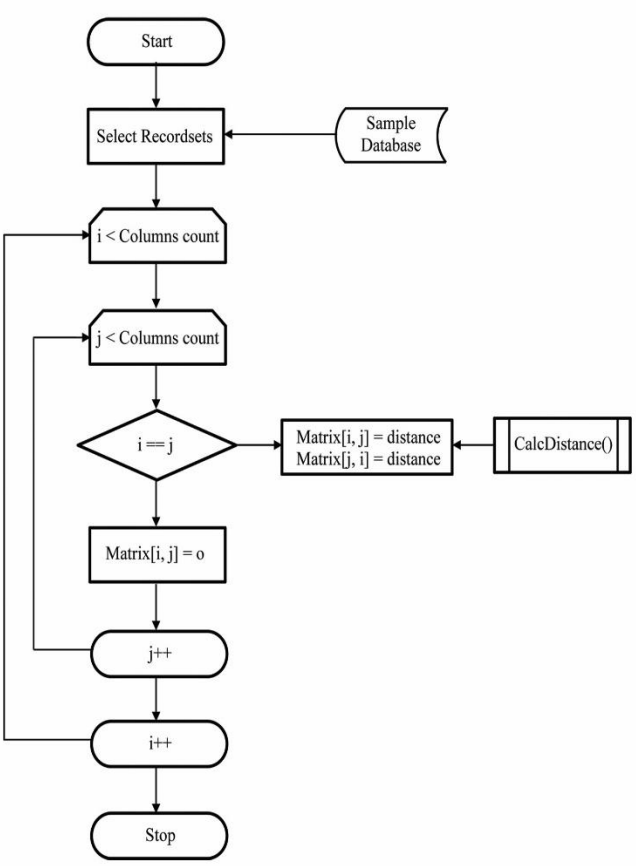


**Fig. 1** Flow Chart for Dissimilarity Matrix Calculation

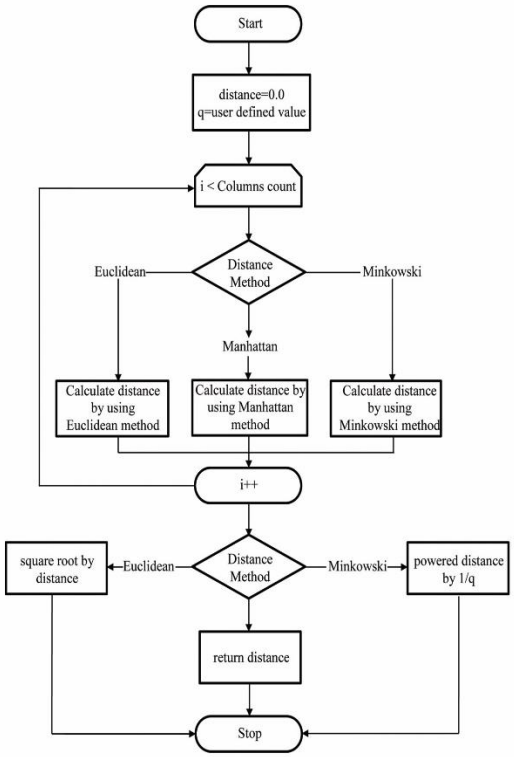


**Fig. 2** Flow Chart for Distance Calculation

## IV. RESULTS

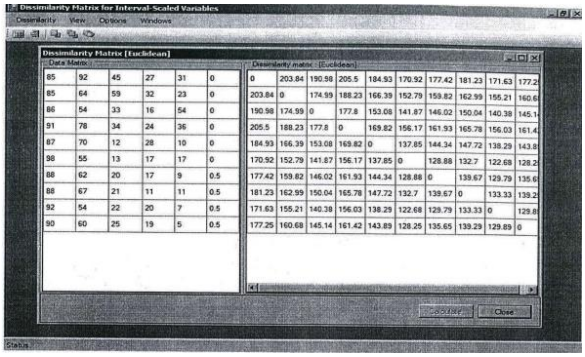

**Fig.3** Dissimilarity Matrix by using Euclidean distance

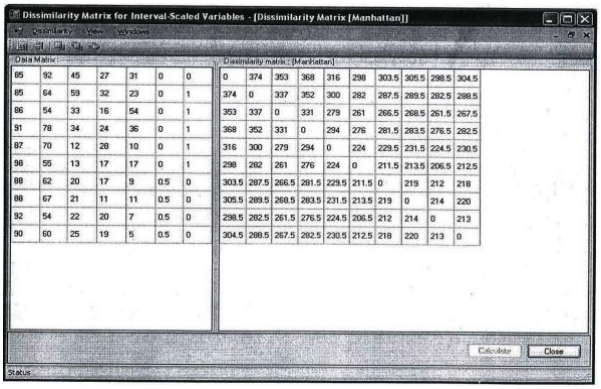

**Fig.4** Dissimilarity Matrix by using Manhattan distance

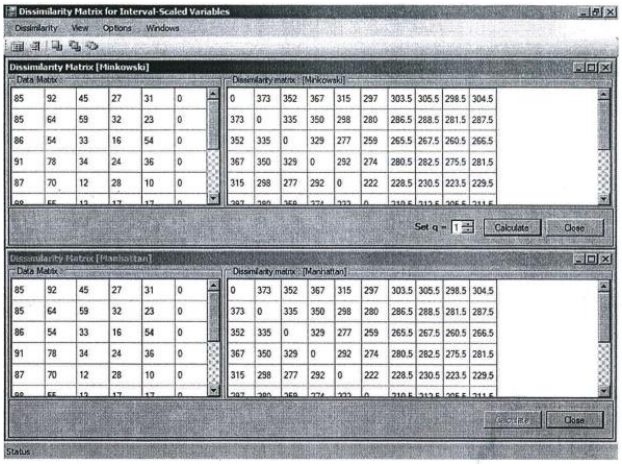

**Fig.5** Dissimilarity Matrix by using Minkowski distance (q=3)

## V. CONCLUSION

A cluster may be a collection of knowledge objects that are almost like each other within an equivalent cluster and are dissimilar to the objects in other clusters. The process of grouping a group of physical or abstract objects into classes of comparable objects is named clustering. Cluster analysis has wide application including market or customer segmentation, pattern recognition, biological studies, spatial data analysis, web document classification, and much of others. Dissimilarity matrix is for memory-based clustering and also called object-by-object structure. Proximities of pairs of objects d (i,j): dissimilarity between objects i and j. In dissimilarity between objects, distances are normally used measures. It then described distance measures that are commonly used for computing the dissimilarity of objects described by liver Disorder variables. These measures include the Euclidean, Manhattan, and Minkowski distances. Minkowski distance is a generalization: if q=2, d is Euclidean distance and if q=1, d is Manhattan distance.

## V. REFERENCES

[1]. Jiawei Han and Micheline Kamber "Data Mining Concepts and Techniques", Morgan Kaufmann, 2001.

[2]. Keim D.A, "Knowledge Discovery and Data Mining", Newport Beach USA, 1997.
[3]. Lu H., Setino R., and Liu H, "Neurorule: A connectionist approach to data mining", VLDB, Switzerland, 1995.
[4]. Pang-Ning, Tan Michael Steinbach and Vipin Kumar, "Introduction to Data Mining".
[5]. Tom M. Mitchell, "Machine Learning", McGraw Hill, New York, 1997

**Cite this article as :**

Aye Aye Thant, Soe Moe Aye, "Euclidean, Manhattan and Minkowski Distance Methods For Clustering Algorithms", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 7 Issue 3, pp. 553-559, May-June 2020. Available at
doi : https://doi.org/10.32628/IJSRSET2073118
Journal URL : http://ijsrset.com/IJSRSET2073118