RAME PUBLISHERS

INTERNATIONAL JOURNAL OF COMPUTATIONAL AND ELECTRONICS ASPECTS IN ENGINEERING

ISSN: 2395-4116 [PRINT]
ISSN: 2395-4124 [ONLINE]

A Quarterly Publishing Journal

International Journal of Computational and Electronics Aspects in Engineering
Vol. 1, Issue 1, Dec 2014, Revised in March 2020, pp 11-15.
https://doi.org/10.26706/ijceae.1.1.20141203

Comparative Study of Euclidean and City Block Distances in Fuzzy C-Means Clustering Algorithm

Saratha Sathasivam[1]
saratha@usm.my

Abdu Masanawa Sagir[2]
ams13_mah013@student.usm.my

University Sains Malaysia,
School of Mathematical
Sciences, Pinang, Malaysia.

***Abstract*— Fuzzy c-means algorithm is one of the most important partitioning techniques and widely used for data clustering and image segmentation. The choice of distance metrics has played key role in data clustering problems since distance metric is used to determine the similarities between data points. In this paper Fuzzy c-means algorithms uses Euclidean and City block distances for comparative analysis to measure the similarities between objects. The results for data clustering problems using Euclidean distance has shown good performance than City block distance in terms of computational time values and the quality of clusters obtained. Similarities, differences and applications of the two proposed distance metrics have been described.**

***Index Terms*—** City block distance, Clustering, Euclidean distance, Fuzzy c-Means

I. Introduction

Clustering is the process of grouping data elements into classes or clusters so that items in each class or cluster are as similar as possible. It is an unsupervised classification designed to group a set of data samples with similar characteristics into a layer units of analysis (clusters). Due to its very significant important in various applications, clustering techniques can be generally classified into different types such include partitional technique, spectral, density – based algorithm, grid – based and model-based approach. Iteratively Clustering algorithm calculates the characteristics of each cluster and sections the image by classifying each pixel in the closest cluster according to a distance metric. Grabusts, [1], the important step in clustering is to select a distance metric, which will determine how the similarity of two elements is calculated. In this paper two different types of distance metrics are used for comparison i.e. Euclidean distance and City block distance.

In the literature, Mohammed [2], described a Hybrid Fuzzy data clustering algorithm using different distance metrics. Francisco et al [3], explains partitioning Fuzzy c-means clustering algorithms for interval – valued data based on City block distances. De Carvalho [4], presented a fuzzy clustering algorithm for symbolic interval data based on a single adaptive Euclidean distance, and De Carvalho [5], Batch self-organizing maps based on city-block distances for interval variables.

This paper aims for comparative analysis of Euclidean and City Block distances based on Fuzzy c-means clustering algorithms.

This paper is organized as follows: section 2 explains briefly the Fuzzy c-means clustering algorithm. Section 3 describes the similarities and differences, and advantages and disadvantages of Euclidean and City block distances, this led us to section 4 which analyze the applications of the two distance measures. Section 5 concludes the remarks.

Research Paper
First Online on – 30 Dec 2014, Revised on – 30 March 2020

Cite this article – Saratha Sathasivam and Abdu Masanawa Sagir, "Comparative Study of Euclidean and City Block Distances in Fuzzy C-Means Clustering Algorithm", *International Journal of Computational and Electronics Aspects in Engineering*, RAME Publishers, vol. 1, issue 1, pp. 11-15, 2014, Revised in 2020.
https://doi.org/10.26706/ijceae.1.1.20141203

II. Training Algorithm

The Fuzzy c-means clustering algorithm using Euclidean distance and City block distance is used. Fuzzy c – means clustering algorithm is the most popular method among the fuzzy clustering techniques [6]. This algorithm works by assigning membership to each data point corresponding to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. The more data is close to the cluster center the more is its membership towards the particular cluster center. Given a finite set of data, the algorithm returns a list of c clusters $C = \{c_1, \dots, c_c\}$ and a partition matrix $J = U = U_{ik} \in [0,1], i = 1, \dots, n;\ k = 1, \dots, c,$ where each element U_{ik} tells the degree to which element x belongs to cluster c_j. Let $X = \{x_1, x_2, \dots, x_N\}$ be the data to be clustered by the FCM algorithm into c (2 ≤ c < N) classes, where N is the number of data points in the design set. The Fuzzy c- means aims to minimize an objective function:

$$J(U,V) = \sum_{i=1}^{N}\sum_{j=1}^{c}(U_{ij})^m \left\|x_i - v_j\right\|^2 \qquad (1)$$

U_{ij} is the degree of membership of x_i in the cluster j, v_j is the center of cluster, $\|*\|$ is any norm expressing the similarity between any measured data (e.g. Euclidean and City block distances) and the center, m is any real number > 1 ($1 \le m < \infty$).

A. The FCM Clustering Algorithm

According to Rabunal [7], the Fuzzy c – means clustering algorithm reads as follows:

1. Initialize the membership values U_{ik} of the x_k, k objects to each of the i clusters, for i = 1,…, c and k = 1, …, n such that:

$$\sum_{i=1}^{c} U_{ik} = 1, \forall k = 1, \dots, n, \forall i = 1, \dots, c \qquad (2)$$

$$and\ U_{ik} \in [0,1] \quad \forall k = 1, \dots, n$$

2. Calculate the cluster center using these membership values:

$$V_i = \frac{\sum_{j=1}^{n}(U_{ik})^m x_k}{\sum_{j=1}^{n}(U_{ik})^m} \ \forall i = 1, \dots, c \qquad (3)$$

3. Calculate the distance metric $D_{[c,n]}$

$$D_{i,j} = \left(\sum_{j=1}^{m}\left\|x_i - v_j\right\|^2\right) \qquad (4)$$

Where $\|*\|$ is any norm expressing the similarity between ith data and jth cluster center.

4. Calculate the new membership values $U_{ik}{}^{new}$ using these cluster centers

$$U_{ik}{}^{new} = \left[\sum_{k=1}^{c}\left[\frac{\left\|x_k - v_i\right\|}{\left\|x_k - v_j\right\|}\right]^{\frac{2}{m-1}}\right]^{-1} \qquad (5)$$

$$\forall k = 1, \dots, n, \forall i = 1, \dots, c$$

5. If $\|\mu^{new} - \mu\| < \varepsilon$, then go to stop otherwise return to step 2 by updating the cluster centers iteratively and also membership grades for the data. Where ε is a termination criterion between 0 & 1, while k is the iterative steps. The exponent m determines the degree of fuzziness of the resulting clustering process. As $m \to 1$ the fuzziness of clustering result tends to the derived with the classical clustering method. As $m \to \infty$ the membership values of all the objects to each cluster tend to the reciprocal of the number of classes $\frac{1}{c}$.

III. Distance Metric

Distance measures can be categorized as metric, semi-metric or non-metric. Clustering methods use distance metrics to determine the similarity or dissimilarity between any pair of objects. The distance between data and centroid can be measured using distance metrics.

According to Hasnat et al [8], a distance is a function δ with non-negative real values, defined on the Cartesian product X x X of a set X. It is call a metric on X if for every x, y, z ∈ X; the following are important properties of distance metric such as:

- $\delta(x,y) = 0, iff\ x = y$ (the identity axiom)

- $\delta(x,y) + \delta(y,z) \geq \delta(x,z)$ (the triangle inequality)
- $\delta(x,y) = \delta(y,x)$ (the symmetry axiom)

Many different measures have been proposed to compute distances between points. These measures are commonly used for algorithms such as clustering, segmentation e.t.c. However, there are similarities and differences, and merits and demerits between distance measures.

A. Similarities and Differences between Euclidean and City block distances.

TABLE I

SIMILARITIES

▪ Euclidean distance and City distance are both distance metrics commonly used for clustering applications
▪ The Euclidean and city-block distances are special cases of the Lp distance.
▪ Euclidean and City block distances are Distance functions for numeric attributes
▪ One of the reasons for using Euclidean and City block distances is the relative ease of their implementation

TABLE II

DISSIMILARITIES

Euclidean distance is usually called Pythagorean metric or L_2 distance	City block distance is usually called Manhattan distance or L_1 distance
The distance between two points x, y in Euclidean n-space is given by $\delta(x,y) = \delta(y,x) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$ x & y are Euclidean vectors	The distance between two points x, y with n-dimension in City block distance is given by $\delta_{xy} = \sum_{i=1}^{n}\lvert x_i - y_i \rvert$
The Euclidean distance is a special case of Minkowski metric when p = 2	The City block distance is a special case of Minkowski metric when p = 1
$\delta_{xy} = \left(\sum_{i=1}^{n}\lvert x_i - y_i\rvert^{\frac{1}{p}}\right)^p$	$\delta_{xy} = \left(\sum_{i=1}^{n}\lvert x_i - y_i\rvert^{\frac{1}{p}}\right)^p$
For normalization process, Normalized Euclidean distance is given by $d = \frac{d_E}{\sqrt{n}}$ d_E is the Euclidean distance.	For normalization process, Normalized City block distance is given by $d = \frac{d_C}{n}$ d_C is the city block distance
Euclidean distance is widely used in distance analyses but it tends to underestimate road distance and travel time	City block distance, on the contrary, tends to overestimate road distance and travel time.

B. Advantages and Disadvantages of the Two Distance Metrics

TABLE III

ADVANTAGES AND DISADVANTAGES OF EUCLIDEAN DISTANCE

Advantages	Disadvantages
The FCM algorithm uses Euclidean distance measure to produce suitable cluster with a spherical shape	Euclidean distance is not suitable for ordinal data, where preferences are listed according to rank instead of according to actual values.
Euclidean distance compares the relationship between actual ratings. That is to say, the Euclidean distance is a fair measure of how similar ratings are for specific preferences or items.	The Euclidean distance suffer establish any correlation where there is a high noise-to-signal ratio and negative spikes.
It is apparently faster than most other means of determining clustering.	Euclidean distance cannot determine the correlation between user profiles that have similar trend in taste, but different ratings for some of the same items

TABLE IV

ADVANTAGES AND DISADVANTAGES OF CITY BLOCK DISTANCE

Advantage	Disadvantage
The FCM algorithm uses City block distance measure to produce suitable clustering for pattern recognition and image processing.	In City block distance, one can only move along one dimension of the space at a time.
City block distance decomposes into contributions made by each variable for the L_2 Euclidean distance.	It is not compatible with many standard multivariate analyses, for example discriminant analysis, Canonical correlation, and canonical corresponding analysis.
	It is more problematic to design algorithms implementing actual road network distance in spatial analytical models.

IV. APPLICATIONS OF THE TWO DISTANCE METRICS

According to [9] and [10], highlighted some of the applications of the two-distance metrics in real life purposes, such include:

- The Euclidean and City distances are applicable in digital image processing (e.g. blurring effects, skeletonizing), motion planning in robotics and even pathfinding.
- Both Euclidean and City distances can be applied in Multidimensional scaling (MDS), which is a technique that uses similarity judgments (or some other proximity measure) to produce a psychological space in which similarity is inversely related to distance.
- In chess, the distance between squares on the chessboard for rooks is measured in City block distance, and bishops use the City block distance (between squares of the same color) on the chessboard rotated 45 degrees, i.e., with its diagonals as coordinate axes.
- City block distance is applicable in Biometric cryptographic.
- Both Euclidean and City block distances could be used in spatial analytical modeling for health service planning.

V. CONCLUSIONS

Traditionally, Euclidean distance is used in clustering algorithm. The results obtained by applying different distance metrics are similar in some cases, that is, none of the metrics shows dominance that allows considering it as the best metric. The choice of distance metric depends on the task, the amount of data and on the complexity of the task. The used of benchmark datasets for two different distance metrics Euclidean and City block based on hybridization techniques have shown that the results using Euclidean distance performed better than city block distance. Conclusively, we observed that the Euclidean distance has shown good performance than City block distance.

Future work: Fuzzy reasoning algorithms should be developed for comparative analysis using four different distance metrics: City block, Chebyshave, Euclidean and Minkowski distances.

ACKNOWLEDGMENT

This research was supported by FRGS grant (203/PMATHS/6711368) offered by Ministry of Higher Education and Universiti Sains Malaysia.

REFERENCES

[1] Grabusts, P. The choice of metrics for clustering algorithms. In *Proceedings of the 8th International Scientific and Practical Conference* (Vol. 2). *ISSN 1691-5402 ISBN 978-9984-44-071-2* Pp. 70 -76, 2011.

[2] Jafar, O. M., & Sivakumar, R. Hybrid Fuzzy Data Clustering Algorithm Using Different Distance Metrics: A Comparative Study. International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume 3, Issue 6, pp. 241 – 248, January 2014.

[3] De Carvalho, F. D. A., & Tenório, C. P. Partitioning Fuzzy c-means clustering for interval-valued data based on adaptive quadratic distances. *Fuzzy Sets and Systems*, *161*(23), 2978-2999, 2010.

[4] De Carvalho, F. D. A. A fuzzy clustering algorithm for symbolic interval data based on a single adaptive Euclidean distance. In *Neural Information Processing*. Springer Berlin Heidelberg. , pp. 1012-1021, January, 2006.

[5] De Carvalho, F. D. A., Bertrand, P., & De Melo, F. M.. Batch self-organizing maps based on city-block distances for interval variables.Version 1, pp1-15, 2012. http://hal.archives-ouvertes.fr/hal-00706519

[6] Lu, Y., Ma, T., Yin, C., Xie, X., Tian, W., & Zhong, S.. Implementation of the Fuzzy C-Means Clustering Algorithm in Meteorological Data. *International Journal of Database Theory and Application,* Vol.6, No.6. Pp.1-18, 2013. http://dx.doi.org/10.14257/ijdta.2013.6.6.01

[7] Ranunal, J. R., & Dorado, J. Artificial Neural Networks in Real-Life Applications. Idea Group Publishing, Hershey, USA. ISBN: 1-59140-902-0, 2006.

[8] Hasnat, A., Halder, S., Bhattacharjee, D., Nasipuri, M., & Basu, D. K.. Comparative Study of Distance Metrics for Finding Skin Color Similarity of Two Color Facial Images. *Computer Science & Information Technology,* pp. 99-108, 2013.

[9] Kimmel, R., Shaked, D., *Kiryati, N., and Bruckstein, A. M., 1996. Online @*
http://en.wikipedia.org/wiki/Distance_transform

[10] Eugene, F. K, 1987. Online @
http://en.wikipedia.org/wiki/Taxicab_geometry