Clustering by Partitioning around Medoids using Distance-Based Similarity Measures on Interval-Scaled Variables

D. L. Nkweteyim

University of Buea, P.O. Box 63, Buea, Cameroon.

ABSTRACT: It is reported in this paper, the results of a study of the partitioning around medoids (PAM) clustering algorithm applied to four datasets, both standardized and not, and of varying sizes and numbers of clusters. The angular distance proximity measure in addition to the two more traditional proximity measures, namely the Euclidean distance and Manhattan distance, was used to compute object-object similarity. The data used in the study comprise three widely available datasets, and one that was constructed from publicly available climate data. Results replicate some of the well known facts about the PAM algorithm, namely that the quality of the clusters generated tend to be much better for small datasets, that the silhouette value is a good, even if not perfect, guide for the optimal number of clusters to generate, and that human intervention is required to interpret generated clusters. Additionally, results also indicate that the angular distance measure, which traditionally has not been widely used in clustering, outperforms both the Euclidean and Manhattan distance metrics in certain situations.

KEYWORDS: PAM, Euclidean, Manhattan, Angular distance, Silhouette.

[Received September 27 2017; Revised December 8 2017; Accepted December 20 2017]

I. INTRODUCTION

Cluster analysis (or clustering) is an unsupervised machine learning task used to find structure in unlabelled data. The clustering task groups a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other clusters (Aldenderfer and Blashfield, 1984; Han *et al*, 2006). Objects to be clustered are typically represented as a matrix, with each row of the matrix representing an object as a features vector, and each dimension of the vector representing a feature or variable used to describe the object (Aldenderfer and Blashfield, 1984; Han *et al*, 2006; Kruskal *et al*, 1978; Tversky, 1977). Interpretation of generated clusters often requires human intervention to explain patterns that are common to members of the clusters.

Application areas of clustering are wide-ranging, including the following: biology, in classification of plants and animals; fraud detection in insurance, by identifying groups of customers with unusually high claims; automatic classification of similar web documents; and marketing, by identifying classes of customers with similar buying habits.

Several clustering approaches have been developed to address different types of data. These include: partitioning approaches, hierarchical approaches, density-based methods, grid-based methods, model-based methods, special techniques for clustering high-dimensional data, and constraint-based clustering (Han *et al*, 2006; Yinghua et al, 2016). For the partitioning approaches, given a collection of **n** objects to cluster, **k** (**k** $\leq$ **n**) clusters of the objects are created, with each cluster containing at least one object, and each object belonging to exactly one cluster. Two common heuristics used to determine cluster membership in partitioning algorithms are a centroid-based technique – K-means, where each cluster centre is represented by the mean value of the objects in the cluster, and a representative object-based technique – K-Medoids, where each cluster centre is represented by one of the **n** objects. Partitioning Around Medoids (PAM), is a widely used K-Medoids method.

The PAM algorithm starts by selecting **k** random objects (medoids) as representative of the **k** clusters, and each of the remaining **n-k** objects assigned to one of the **k** clusters based on how similar the objects are to the corresponding medoids. The algorithm then attempts to improve on the initial clustering as follows: for each medoid object, swap each of the non-medoid objects with the medoid and compute the cost, i.e., average dissimilarity, of the new clustering that results from this swap. If the cost increases, then the swap is undone.

Determination of object-object dissimilarity and hence cluster membership of objects has traditionally been done using the Euclidean and Manhattan (or city block) distance measures. Angular distance (the angular separation between two objects), another distance metric, has not been as widely studied in cluster analysis. The objective of this study was to implement and apply the PAM algorithm to both standardized and non-standardized versions of four datasets, and for each version of a dataset, compare the quality of clusters obtained from the Euclidean distance, Manhattan distance and angular distance similarity metrics. In each case, the quality of the generated clusters was determined using two metrics: the proportion of objects placed in the correct cluster as

*Corresponding author's e-mail address: nkweteyim.denis@ubuea.cm

doi: http://dx.doi.org/10.4314/njtd.v15i1.1

suggested by the each object's predetermined class in the dataset, and the average silhouette scores of the clusters.

II. MATERIALS AND METHODS

A. Materials and Datasets

1) Iris dataset: This data set contains 150 instances of the iris plant comprising 50 instances each, of the following three classes of the plant: Iris Setosa, Iris Versicolour, and Iris Virginica. One class (Iris Setosa) is linearly separable from the other two classes, which are not linearly separable from each other. Four attributes are used to describe each plant: sepal length, sepal width, petal length, and petal width, all measured in units of cm. (Linchman, 2013a).

2) Leaf dataset: The dataset comprises data on the leaves of 30 different plant species. There is a total of 339 leaf specimens and for each specimen, there are 7 attributes (eccentricity, aspect ratio, elongation, solidity, stochastic convexity, isoperimetric factor, and maximal indentation depth) describing shape, and another 7 attributes (lobedness, average intensity, average contrast, smoothness, third moment, uniformity, and entropy) describing texture. (Linchman, 2013b; Silva *et al*, 2013).

3) Climate dataset: This dataset comprises climate data on 171 locations in Cameroon with 6 different climate classes based on the Köppen climate classification system (Koppen Climate Classification, n.d.; The Koppen Climate Classification System, n.d.), extracted from the climate-data.org web site (Climate-data.org, n.d.). For each location, there are 48 attributes, namely the average monthly precipitation in mm, the average monthly, minimum monthly, and maximum monthly temperatures in Celsius for each month of the year. The Köppen climate classification system identifies four basic climate zones and ten sub-climate zones as follows:

i.) Cold climate (Taiga Biome – Boreal Forest Climate: Dfc; Tundra Biome – Tundra Climate: E; Alpine Biome – Highland Climate: H)
ii.) Tropical climate (Tropical Rainforest – Tropical Moist Climates: Af; Savanna – Wet-Dry Tropical Climates: Aw; Chaparral Biome – Mediterranean Climate: Cs)
iii.) Dry climate (Desert Biome – Dry Tropical Climate: Bw; Steppe – Dry Midlatitude Climate: BS)
iv.) Temperate climate (Deciduous Forest Biome – Continental Climate: Cf; Grasslands Biome – Dry Midaltitude Climates: Bs)

The 171 climate data points that were used in this work comprise six climate categories: (Af: 10; Am: 56; Aw: 69; Bsh: 4, Bwh: 5; Cwb: 27)

4) Mice protein dataset: This dataset consists of the expression levels of 77 proteins/protein modifications that produced detectable signals in the nuclear fraction of cortex of mice. 15 measurements were made for each of 72 mice (38 control and 34 trisomic, i.e. with Down syndrome) (Higuera *et al*, 2015; Linchman, 2013c). So, altogether, there were 72 x 15, or 1,080 measurements per protein. Each measurement can be considered as an independent sample/mouse. However, there are some missing data in the dataset. In this work, all the samples with missing data were removed, leaving 552 out of the original 1,080 samples.

The mice are placed into eight classes based on features such as genotype (control or trisomic), behavior (context-shock for mice that have been stimulated to learn, and shock-context for mice that have not been stimulated to learn), and treatment using the drug memantine in recovering the ability to learn in trisomic mice (some mice injected with the drug and others not injected). The resulting eight classes were as follows:

i.) c-CS-s: control mice, stimulated to learn, injected with saline (5 mice, 75 measurements)
ii.) c-CS-m: control mice, stimulated to learn, injected with memantine (3 mice, 45 measurements)
iii.) c-SC-s: control mice, not stimulated to learn, injected with saline (5 mice, 75 measurements)
iv.) c-SC-m: control mice, not stimulated to learn, injected with memantine (4 mice, 60 measurements)
v.) t-CS-s: trisomy mice, stimulated to learn, injected with saline (5 mice, 75 measurements)
vi.) t-CS-m: trisomy mice, stimulated to learn, injected with memantine (6 mice, 90 measurements)
vii.) t-SC-s: trisomy mice, not stimulated to learn, injected with saline (5 mice, 72 measurements)
viii.) t-SC-m: trisomy mice, not stimulated to learn, injected with memantine (4 mice, 60 measurements)

B. Theoretical Framework and Methodology

1) Determining an Optimal Value for k, the number of clusters to generate: A common approach makes use of a silhouette (Rousseeuw, 1987; "Silhouette (clustering)," 2016). A silhouette value is a computed number that ranges from -1 to +1, and is a compact measure for cohesion (how similar an object is to its cluster) and separation (how dissimilar an object is to other clusters). A high silhouette value suggests that the object is well matched to its cluster, and a low silhouette value suggests that the object is not well matched to its cluster. Hence, in determining an appropriate value for ***k***, clusters that have many objects with high silhouette values and few objects with low silhouette values were sought. The average silhouette score for each cluster is an indicator of how good the cluster is; similarly, the average silhouette value for all the clusters is an indicator of how good all the clusters put together are.

2) Object Representation: An important consideration in clustering algorithms is the data types of the variables used to represent objects, as these determine the approach to compute similarity between objects. Variable types include the following (Bramer, 2013; Han *et al*, 2006): interval-scaled variables, binary variables, categorical variables, ordinal variables, and ratio-scaled variables.

Interval-scaled variables are continuous measurements on a roughly linear scale (Han *et al*, 2006). Examples are weight, length, and temperature. One concern with interval-scaled variables is that the measuring unit may affect the generated clusters. Expressing a variable in smaller units (for example, centimetres instead of metres) leads to a larger range for that variable, and thus a larger effect on the resulting clustering structure (Han *et al*, 2006). One way to

avoid dependence of the clustering algorithm on the measurement unit is to standardize the data, either by giving each variable the same weight, or in some cases, giving higher weights to important variables (e.g., height of a basketball player when clustering potential recruits for a basketball team). It is important to note though, that standardization may or may not be useful in some applications. This work considers only objects represented with interval-scaled variables, and limit discussion to this category of variables.

3) Similarity Measures: Several proximity measures are available to choose from, including spatial models, set-theoretic models and graph-theoretic models (Corter, 1996); correlation coefficients, distance measures, association coefficients, and probabilistic similarity measures (Aldenderfer and Blashfield, 1984). Distance-based proximity measures consider objects as points in a coordinate space; with such a representation, object-object similarity is a measure of how close the objects are in this space. A true distance-based similarity metric must meet four criteria (Aldenderfer and Blashfield, 1984), summarized below for objects **x**, **y**, and **z**, separated by distance **d**:

i.) *Symmetry*: the direction of measurement of distance is immaterial, i.e., $d(x,y) = d(y,x) \geq 0$

ii.) *Distinguishability of nonidenticals*: if the two points differ, then the distance between them is not equal to zero, and from the symmetry criterion, must be greater than zero. In other words, since $d(x,y) \neq 0$, then $x \neq y$.

iii.) *Indistinguishability of incidentals*: if two points coincide, then the distance bewteen them is zero, i.e., $d(x,x) = 0$

iv.) *Triangle inequality*: for any three points, x, y, z, the following relationship holds true $d(x,y) \leq d(x,z) + d(y,z)$

Two of the most commonly used distance metrics are Euclidean distance and Manhattan distance (Aldenderfer and Blashfield, 1984; Han *et al*, 2006; Tversky, 1977). The cosine similarity measure has been widely used to in areas like information retrieval and text mining that make use of high-dimensional data (Han *et al*, 2006; Manning., 2008; Nkweteyim, 2014; Salton, 1988; Salton and McGill, 1986), and can also be used in cluster analysis.

The cosine similarity metric works well in information retrieval systems using the vector-space model to represent objects. In that representation, document vectors are represented by term weights, which are all positive, guaranteeing that the similarity score always ranges from 0 to 1. However, in other applications in which object variables may be negative, the cosine similarity score ranges from -1 to 1. In such cases, cosine similarity does not meet the triangle inequality requirement for a distance-based similarity metric, and the angular distance metric, which meets this requirement, could be used instead.

Once determined, object-object proximity measures can be stored in a look-up table to be consulted during the clustering process.

4) Standardization of Object Vectors: As mentioned in the introduction, an object can be represented in standard form as a features vector with the weight given to each feature dependent on the unit of measure used. An alternative representation is to standardize the vector to give each variable the same weight. One common way to standardize data is to compute a z-score (Equation 3) for each variable, as illustrated below.

Given $\boldsymbol{p}$ objects each with $\boldsymbol{n}$ variables: $Object1 = x_{11}, x_{12}\ldots, x_{1n}, Object2 = x_{21}, x_{22}\ldots, x_{2n}, \ldots \quad Objectp = x_{p1}, x_{p2}\ldots, x_{pn}$

Now consider variable $\boldsymbol{f}$ for each of the $\boldsymbol{p}$ objects. The z-score Z_{if} for variable $\boldsymbol{f}$ of object $\boldsymbol{i}$ is computed as follows (see Han *et al*, 2006):

Calculate the mean absolute deviation S_f:

$$S_f = \frac{1}{n}(| x_{1f} - m_f | + | x_{2f} - m_f | + \cdots + | x_{pf} - m_f |), \quad (1)$$

where $x_{1f}, x_{2f}, \ldots x_{pf}$ are $\boldsymbol{p}$ measurements of $\boldsymbol{f}$ and $\boldsymbol{m_f}$ is the mean value of $\boldsymbol{f}$, i.e.,

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \ldots + x_{pf}) \quad (2)$$

Calculate the z-score for variable $\boldsymbol{f}$ of objet $\boldsymbol{x_i}$ from Equations 1 and 2:

$$Z_{if} = \frac{x_{if} - m_f}{s_f} \quad (3)$$

5) Similarity Measures used: Given two objects each with $\boldsymbol{n}$ variables, $j = x_{j1}, x_{j2}\ldots, x_{jn}$ and $k = x_{k1}, x_{k2}\ldots, x_{kn}$, the distance between them using four common similarity metrics used in this study are given in Equations 4 – 7 below.

Euclidean distance

$$d(j,k) = \sqrt{(x_{j1} - x_{k1})^2 + (x_{j2} - x_{k2})^2 + \ldots + (x_{jn} - x_{kn1})^2} \quad (4)$$

Manhattan distance

$$d(j,k) = | x_{j1} - x_{k1} | + | x_{j2} - x_{k2} | + \ldots + | x_{jn} - x_{kn1} | \quad (5)$$

Cosine similarity

$$\cos(\theta) = \frac{x_{j1}x_{k1} + x_{j2}x_{k2} + \ldots + x_{jn}x_{kn1}}{\sqrt{{x_{j1}}^2 + {x_{j2}}^2 + \ldots + {x_{jn}}^2}\sqrt{{x_{k1}}^2 + {x_{k2}}^2 + \ldots + {x_{kn}}^2}} \quad (6)$$

Angular $distance = \frac{\arccos(similarity)}{\pi} = \frac{\arccos(\cos(\theta))}{\pi}$ (7)

6) Silhouette values: These can be determined as outlined below:

Consider a cluster **A**, containing an object $\boldsymbol{i}$.

Let $\boldsymbol{a(i)}$ be the average dissimilarity of object $\boldsymbol{i}$ with all other data in cluster **A**

Let $\boldsymbol{b(i)}$ be the lowest average dissimilarity of object $\boldsymbol{i}$ with all other clusters different from cluster **A**. Such a cluster (cluster **B**) is a neighbour to cluster **A**

A silhouette $\boldsymbol{s(i)}$ for object $\boldsymbol{i}$ is defined as: $s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$ (8)

Equation 8 can be rewritten as:

$$\begin{cases} 1 - a(i)/b(i), & if\, a(i) < b(i) \\ 0, & if\, a(i) = b(i) \\ b(i)/a(i) - 1 & if\, a(i) > b(i) \end{cases} \quad (9)$$

It is noticed always that, $-1 \leq s(i) \leq 1$.

III. METHODOLOGY

In this study, C code was designed and implemented to generate clusters using the PAM algorithm as well as compute silhouette values for different numbers of clusters generated from the four datasets, both standardized and not standardized. In each case, the Euclidean, Manhattan, and angular distance were used to determine the dissimilarities between objects. Guided by the number of classes ***k***, of objects that were suggested by the datasets, five sets of clusters were generated for all but the iris dataset comprising k-2, k-1, k, k+1, and k+2 clusters respectively in a bid to appreciate the usefulness of silhouette values in guiding an ideal number of clusters to be generated for a given dataset; for the iris dataset with only three categories of flowers, the number of clusters generated were 2, 3, 4, 5, and 6.

IV. RESULTS AND DISCUSSION

A. Clusters

Tables 1-4 show the clustering results for the four datasets. The results are based on the optimal number of clusters generated for the number of classes of objects suggested by the dataset: 3, 6, 8, and 30 clusters respectively for the iris, climate, mice, and leaf datasets.) The results show that for non-standardized data, the angular distance measure outperforms the Euclidean and Manhattan distance measures for three of the four datasets, while the Manhattan distance measure fares similarly to, or slightly better than, the Euclidean distance measure, on all the non-standardized datasets. When the datasets are standardized however, there is no clear advantage of any of the three measures over the others. These facts are further summarized and illustrated in Figures 1 and 2.

Table 1: Percentages of correct classification into three clusters generated from the iris dataset using different similarity measures.

Dataset details		**Dataset not standardized**			**Dataset standardized**		
Class	Count	Euclid (%)	Manhattan (%)	Angular Distance (%)	Euclid (%)	Manhattan (%)	Angular Distance (%)
Iris-setosa	50	100.0	100.0	100.0	100.0	100.0	98.0
Iris-versicolor	50	96.0	92.0	90.0	80.0	86.0	62.0
Iris-virginica	50	72.0	72.0	100.0	70.0	76.0	84.0
Total	150	89.3	88.0	96.7	83.3	87.3	81.3

Table 2: Percentages of correct classification into six clusters generated from the climate dataset using different similarity measures.

Dataset details		**Dataset not standardized**			**Dataset standardized**		
Class	Count	Euclid (%)	Manhattan (%)	Angular Distance (%)	Euclid (%)	Manhattan (%)	Angular Distance (%)
Af	10	50.0	50.0	50.0	90.0	90.0	50.0
Am	56	39.3	37.5	67.9	44.6	41.1	39.3
Aw	69	29.0	31.9	56.5	33.3	31.9	34.8
Bsh	4	100.0	100.0	75.0	100.0	100.0	100.0
Bwh	5	100.0	100.0	100.0	100.0	100.0	100.0
Cwb	27	55.6	55.6	100.0	100.0	100.0	100.0
Total	171	41.5	42.1	68.4	54.4	52.6	50.9

Table 3: Percentages of correct classification into eight clusters generated from the mice protein dataset using the similarity measures.

Dataset details		**Dataset not standardized**			**Dataset standardized**		
Class	Count	Euclid (%)	Manhattan (%)	Angular Distance (%)	Euclid (%)	Manhattan (%)	Angular Distance (%)
c-CS-m	45	26.7	31.1	60.0	33.3	33.3	40.0
c-SC-m	60	55.0	53.3	73.3	41.7	51.7	50.0
c-CS-s	75	41.3	48.0	50.7	48.0	44.0	40.0
c-SC-s	75	66.7	69.3	88.0	68.0	66.7	60.0
t-CS-m	90	28.9	38.9	30.0	43.3	47.8	40.0
t-SC-m	60	35S.0	35.0	60.0	41.7	46.7	26.7
t-CS-s	75	46.7	46.7	54.7	30.7	33.3	37.3
t-SC-s	72	34.7	29.2	51.4	36.1	40.3	51.4
Total	552	42.2	44.6	57.2	43.5	46.0	43.5

Table 4: Percentages of correct classification into thirty clusters generated from the leaf dataset using the Euclidean, Manhattan, and Angular Distance similarity measures.

Dataset details		**Dataset not standardized**			**Dataset standardized**		
Class	Count	Euclid (%)	Manhattan (%)	Angular Distance (%)	Euclid (%)	Manhattan (%)	Angular Distance (%)
Quercus suber	12	58.3	58.3	66.7	50.0	58.3	66.7
Salix atrocinera	10	50.0	50.0	50.0	90.0	90.0	90.0
Populus nigra	10	40.0	60.0	60.0	70.0	70.0	60.0
Alnus sp.	8	25.0	37.5	37.5	37.5	37.5	37.5
Quercus robur	12	75.0	91.7	50.0	58.3	58.3	58.3
Crataegus monogyna	8	62.5	75.0	75.0	75.0	87.5	75.0
Ilex aquifolium	10	30.0	50.0	60.0	50.0	50.0	40.0
Nerium oleander	11	100.0	100.0	100.0	100.0	100.0	100.0
Betula pubescens	14	42.9	50.0	28.6	35.7	28.6	42.9
Tilia tomentosa	13	69.2	69.2	61.5	53.8	46.2	53.8
Acer palmatum	16	81.3	87.5	87.5	93.8	93.8	87.5
Celtis sp.	12	50.0	41.7	41.7	33.3	50.0	41.7
Corylus avellana	13	61.5	38.5	38.5	46.2	46.2	46.2
Castanea sativa	12	66.7	75.0	58.3	41.7	41.7	41.7
Populus alba	10	70.0	90.0	70.0	70.0	70.0	80.0
Primula vulgaris	12	41.7	33.3	33.3	50.0	41.7	41.7
Erodium sp.	11	63.6	63.6	63.6	63.6	63.6	54.5
Bougainvillea sp.	13	30.8	30.8	30.8	38.5	61.5	46.2
Arisarum vulgare	9	88.9	88.9	77.8	66.7	66.7	77.8
Euonymus japonicus	12	58.3	50.0	41.7	25.0	16.7	25.0
Ilex perado ssp. Azorica	11	54.5	45.5	54.5	63.6	45.5	45.5
Magnolia soulangeana	12	50.0	50.0	41.7	41.7	41.7	41.7
Buxus sempervirens	12	91.7	91.7	91.7	100.0	100.0	100.0
Urtica dioica	12	41.7	50.0	50.0	50.0	50.0	41.7
Podocarpus sp.	11	36.4	36.4	63.6	54.5	54.5	54.5
Acca sellowiana	11	63.6	72.7	54.5	54.5	45.5	45.5
Hydrangea sp.	11	54.5	54.5	36.4	63.6	54.5	63.6
Pseudosasa japonica	11	54.5	72.7	100.0	72.7	72.7	100.0
Magnolia grandi Ora	11	72.7	81.8	72.7	54.5	54.5	72.7
Geranium sp.	10	10.0	10.0	30.0	20.0	10.0	20.0
Total	340	57.1	60.3	57.4	57.4	56.8	58.2

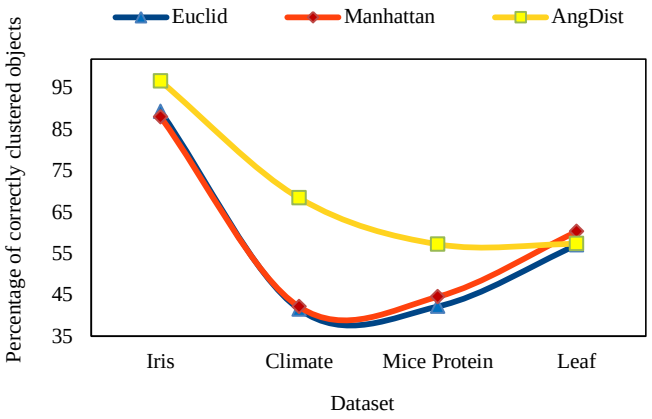

Figure 1: Percentage of objects per dataset correctly clustered on un-normalized datasets.

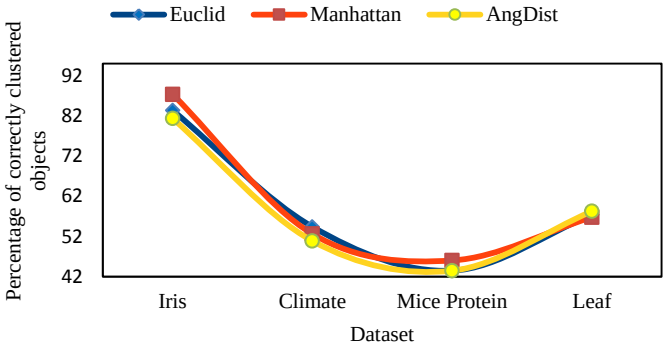

Figure 2: Percentage of objects per dataset correctly clustered on normalized datasets.

B. Silhouette widths

Table 5 shows the average silhouette widths for five consecutive cluster sizes generated from the clustering algorithm, with the silhouette values highlighted for the optimal number of clusters as suggested by the number of classes in the dataset. The optimal numbers of classes as suggested by the respective datasets are printed in bold. With the exception of the iris dataset for which there is a sharp change in silhouette values from one run to another, the changes for the other datasets are mild.

Secondly, the computed silhouette widths range from 0.16 (using Manhattan distance dissimilarity measure on the generation of 8 clusters from the standardized mice protein dataset), to 0.7 (using the angular distance dissimilarity measure on the generation of 6 clusters from the non-standardized climate dataset).

Results suggest that angular distance, which in the past has not been prominently highlighted as a useful dissimilarity metric in clustering, does indeed result in higher quality clusters than the well-known Euclidean and Manhattan distance metrics, in some situations.

Table 5: Average silhouette widths for various runs of the clustering algorithms.

	Dataset not standardized				**Dataset standardized**		
Iris dataset							
	Euclidean	Manhattan	Angular Distance	No. of clusters	Euclidean	Manhattan	Angular Distance
2 clusters	0.72	0.70	0.80	2	0.60	0.65	0.64
3 clusters	**0.56**	**0.56**	**0.56**	**3**	**0.46**	**0.49**	**0.59**
4 clusters	0.51	0.48	0.44	4	0.44	0.44	0.52
5 clusters	0.47	0.49	0.44	5	0.41	0.40	0.46
6 clusters	0.40	0.42	0.33	6	0.40	0.39	0.44
Climate dataset							
	Euclidean	Manhattan	Angular Distance	No. of clusters	Euclidean	Manhattan	Angular Distance
4 clusters	0.49	0.54	0.69	4	0.48	0.53	0.57
5 clusters	0.45	0.47	0.75	5	0.47	0.51	0.51
6 clusters	**0.50**	**0.50**	**0.70**	**6**	**0.47**	**0.49**	**0.50**
7 clusters	0.57	0.57	0.63	7	0.52	0.55	0.46
8 clusters	0.63	0.62	0.58	8	0.57	0.52	0.52
Mice Protein dataset							
	Euclidean	Manhattan	Angular Distance	No. of clusters	Euclidean	Manhattan	Angular Distance
6 clusters	0.21	0.19	0.22	6	0.20	0.14	0.17
7 clusters	0.22	0.20	0.22	7	0.15	0.16	0.18
8 clusters	**0.22**	**0.20**	**0.23**	**8**	**0.17**	**0.16**	**0.17**
9 clusters	0.21	0.19	0.23	9	0.20	0.17	0.18
10 clusters	0.23	0.21	0.23	10	0.21	0.17	0.22
Leaf dataset							
	Euclidean	Manhattan	Angular Distance	No. of clusters	Euclidean	Manhattan	Angular Distance
28 clusters	0.49	0.48	0.42	28	0.37	0.36	0.38
29 clusters	0.48	0.45	0.41	29	0.37	0.36	0.38
30 clusters	**0.50**	**0.44**	**0.43**	**30**	**0.38**	**0.36**	**0.38**
31 clusters	0.50	0.47	0.43	31	0.37	0.36	0.37
32 clusters	0.49	0.45	0.43	32	0.39	0.37	0.37

This is the case with the unstandardized *iris*, *climate*, and *mice protein* datasets, although not so for the corresponding standardized datasets in which angular distance is only superior on the *leaf* dataset.

The second major observation is that silhouette widths are not a silver bullet in determining the number of clusters to

generate using a partitioning clustering algorithm as there is no fine line separating the silhouhette values between the optimal number of clusters and a less optimal number. In this work, the advantage of the pre-determined number of classes in each dataset to know the optimal number of clusters to generate was taken. In practice though, it is not known beforehand, how many clusters to generate. This work thus re-emphasizes the need for human intervention in interpreting generated clusters as silhouette values alone can only serve as a guide to the number of clusters to generate.

Results reveal that the algorithm succeeded in correctly clustering larger percentages of objects in some datasets than others. Too much cannot be read into this as several factors could be responsible, notably variation in dataset sizes, number of attributes, and number of classes. In general, the larger these values are, the more difficult the clustering problem is. The iris dataset with the best performance for example, was the simplest, with 150 objects, 3 clusters, and 4 attributes. The mice protein dataset on the other hand, which performed poorly, comprised 552 objects, eight clusters, and 77 attributes.

There are some limitations in the work, which a similar study could consider, to get a better appreciation of the relative performances of different dissimilarity metrics. First, datasets with similar characteristics (size, number of attributes and number of known classes) could be selected.

With a variety of similarity metrics available for use in clustering, with none of them apparently outperforming all others in all situations, an approach to clustering can be envisaged in which cluster membership of an object is determined not only from the object-object similarity score of a single similarity metric, but rather through a voting system in which a majority of the similarity metrics used, supports membership of the object in that cluster.

V. CONCLUSION

Presented in this paper, were details of the clusters obtained when the PAM algorithm was applied to four datasets (both un-standardized and standardized using the z-score), and using three metrics – Euclidean distance, Manhattan distance, and angular distance – to determine similarity between objects, and hence cluster membership of objects. Cluster silhouette widths were also computed in a bid to appreciate the usefulness of this metric in the estimation of the quality of generated clusters.

Results show that the seldom-used angular distance metric outperforms the widely Euclidean and Manhattan distance dissimilarity metrics in certain situations, and so should be considered as a viable, alternative distance measure by researchers in the area of clustering. Given that different proximity measures result in different clusters, perhaps automated distance-based clustering should use several proximity measures, and place an object in a cluster only if the majority of the proximity measures *vote* for the object to be placed in that cluster.

The work also confirms the fact that silhouette widths alone are not sufficient to determine the quality of generated clusters, and so human examination remains important in interpreting generated clusters.

Low-dimensional datasets was chosen in the work. This work can be extended to investigate other similarity measures to appreciate their usefulness in the PAM algorithm, as well as higher-dimensional data, in order to investigate the effects of dimensionality on the algorithm.

REFERENCES

Aldenderfer, M.S. and Blashfield, R.K. (1984). Cluster Analysis. Sage Publications, United States.

Bramer, M. (2013). Principles of Data Mining: Undergraduate Topics in Computer Science, Springer-Verlag, London.

Climate-data.org (n.d.). Climate: Cameroon. Available at: http://en.climate-data.org/country/142/. Accessed on July 30, 2016.

Han, J.; J. Pei and M. Kamber. (2006). Data Mining: Concepts and techniques, Southeast Asia Edition, 2nd edition. Morgan Kaufmann, Elsevier, Singapore.

Higuera, C., Gardiner, K., Cios, K. (2015). Self-Organizing Feature Maps Identify Proteins Critical to Learning in a Mouse Model of Down Syndrome. PLoS ONE 10(6): e0129126.

Koppen Climate Classification (N.D.). Koppen Climate Classification. Available at: https://www.britannica.com/science/Koppen-climate-classification. Accessed on 3 August, 2016.

Kruskal, J. and Wish, M. (1978). Multidimensional Scaling. Sage Publications, United States.

Linchman, M. (2013a). UCI Machine Learning Repository: Iris Data Set. University of California, School of Information and Computer Science, Irvine, CA.

Linchman, M. (2013b). UCI Machine Learning Repository: Leaf Data Set. University of California, School of Information and Computer Science, Irvine, CA.

Linchman, M. (2013c). UCI Machine Learning Repository: Mice Protein Expression Data Set. University of California, School of Information and Computer Science, Irvine, CA.

Silva, P.F.B.; A.R.S. Marçal and R.M.A. da Silva. (2013). Evaluation of Features for Leaf Discrimination, in: Kamel, M., Campilho, A. (Eds.), Image Analysis and Recognition: 10th International Conference, ICIAR 2013, Póvoa Do Varzim, Portugal, June 26-28, 2013. Proceedings. Springer Berlin Heidelberg, Berlin, Heidelberg, 197–204.

The Köppen Climate Classification System (n.d.). The Koppen Climate Classification System. Available at: http://www.thesustainabilitycouncil.org/resources/the-koppen-climate-classification-system/. Accessed Aug 3, 2016.

Tversky, A. (1977). Features of similarity. Psychological Review, 84: 327–352.

Yinghua, L.; M. Tinghuai, T. Meili, C. Jie, T. Yuan, A. Abdullah and A. Mznah. (2016). An Efficient and Scalable Density-based Clustering Algorithm for Datasets with Complex Structures. Neurocomputing, 171: 9-22.