RESEARCH ARTICLE

A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data

Ali Seyed Shirkhorshidi[1]*, Saeed Aghabozorgi[2], Teh Ying Wah[1]

1 Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, 50603, Kuala Lumpur, Malaysia, **2** IBM Analytics, Platform, Emerging Technologies, IBM Canada Ltd., Markham, Ontario L6F 1C7, Canada

* shirkhorshidi_ali@yahoo.co.uk

OPEN ACCESS

Citation: Shirkhorshidi AS, Aghabozorgi S, Wah TY (2015) A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data. PLoS ONE 10(12): e0144059. doi:10.1371/journal.pone.0144059

Editor: Andrew R. Dalby, University of Westminster, UNITED KINGDOM

Received: May 10, 2015

Accepted: November 12, 2015

Published: December 11, 2015

Data Availability Statement: All third-party datasets used in this study are available publicly in UCI machine learning repository: http://archive.ics.uci.edu/ml and Speech and Image Processing Unit, University of Eastern Finland: http://cs.joensuu.fi/sipu/datasets/ **References are mentioned in the manuscript in "experimental result" and "acknowledgment" sections.

Funding: This work is supported by University of Malaya Research Grant no vote RP028C-14AET. IBM Canada Ltd funder provided support in the form of salaries for author [SA], but did not have any additional role in the study design, data collection and

Abstract

Similarity or distance measures are core components used by distance-based clustering algorithms to cluster similar data points into the same clusters, while dissimilar or distant data points are placed into different clusters. The performance of similarity measures is mostly addressed in two or three-dimensional spaces, beyond which, to the best of our knowledge, there is no empirical study that has revealed the behavior of similarity measures when dealing with high-dimensional datasets. To fill this gap, a technical framework is proposed in this study to analyze, compare and benchmark the influence of different similarity measures on the results of distance-based clustering algorithms. For reproducibility purposes, fifteen publicly available datasets were used for this study, and consequently, future distance measures can be evaluated and compared with the results of the measures discussed in this work. These datasets were classified as low and high-dimensional categories to study the performance of each measure against each category. This research should help the research community to identify suitable distance measures for datasets and also to facilitate a comparison and evaluation of the newly proposed similarity or distance measures with traditional ones.

Introduction

One of the biggest challenges of this decade is with databases having a variety of data types. Variety is among the key notion in the emerging concept of big data, which is known by the 4 Vs: Volume, Velocity, Variety and Variability [1,2]. Currently, there are a variety of data types available in databases, including: interval-scaled variables (salary, height), binary variables (gender), categorical variables (religion: Jewish, Muslim, Christian, etc.) and mixed type variables (multiple attributes with various types). Despite data type, the distance measure is a main component of distance-based clustering algorithms. Partitioning algorithms, such as k-means, k-medoids and more recently soft clustering approaches for instance fuzzy c-means [3] and rough clustering [4], are mainly dependent on distance measures to recognize clusters in a dataset.

In data mining, ample techniques use distance measures to some extent. Clustering is a well-known technique for knowledge discovery in various scientific areas, such as medical

analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Competing Interests: The authors have the following interests: Saeed Aghabozorgi is employed by IBM Canada Ltd. There are no patents, products in development or marketed products to declare. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials, as detailed online in the guide for authors.

image analysis [5–7], clustering gene expression data [8–10], investigating and analyzing air pollution data [11–13], power consumption analysis [14–16], and many more fields of study. Improving clustering performance has always been a target for researchers. Since in distance-based clustering similarity or dissimilarity (distance) measures are the core algorithm components, their efficiency directly influences the performance of clustering algorithms. These algorithms use similarity or distance measures to cluster similar data points into the same clusters, while dissimilar or distant data points are placed into different clusters. Examples of distance-based clustering algorithms include partitioning clustering algorithms, such as k-means as well as k-medoids and hierarchical clustering [17].

Although there are various studies available for comparing similarity/distance measures for clustering numerical data, but there are two difference between this study and other existing studies and related works: first, the aim in this study is to investigate the similarity/distance measures against low dimensional and high dimensional datasets and we wanted to analyse their behaviour in this context. Second thing that distinguish our study from others is that our datasets are coming from a variety of applications and domains while other works confined with a specific domain. In essence, the target of this research is to compare and benchmark similarity and distance measures for clustering continuous data to examine their performance while they are applied to low and high-dimensional datasets. For the sake of reproducibility, fifteen publicly available datasets [18,19] were used for this study, so future distance measures could consequently be evaluated and compared with the results of traditional measures discussed in this study. These datasets are classified into low and high-dimensional, and each measure is studied against each category. But before doing the study on similarity or dissimilarity measures, it needs to be clarified that they have significant influence on clustering quality and are worthwhile to be studied. In sections 3 (methodology) it is elaborated that the similarity or distance measures have significant influence on clustering results.

The key contributions of this paper are as follows:

- Twelve similarity measures frequently used for clustering continuous data from various fields are compiled in this study to be evaluated in a single framework. Most of these similarity measures have not been examined in domains other than the originally proposed one.
- A technical framework is proposed in this study to analyze, compare and benchmark the influence of different similarity measures on the result of distance-based clustering algorithms.
- Similarity measures are evaluated on a wide variety of publicly available datasets. Particularly, we evaluate and compare the performance of similarity measures for continuous data against datasets with low and high dimension.

The rest of paper is organized as follows: in section 2, a background on distance measures is discussed. In section 3, we have explained the methodology of the study. Experimental results with a discussion are represented in section 4, and section 5 summarizes the contributions of this study.

Background on Distance Measures for Continuous Data

Utilization of similarity measures is not limited to clustering, but in fact plenty of data mining algorithms use similarity measures to some extent. To reveal the influence of various distance measures on data mining, researchers have done experimental studies in various fields and have compared and evaluated the results generated by different distance measures. Although it is not practical to introduce a "Best" similarity measure or a best performing measure in general, a comparison study could shed a light on the performance and behavior of measures. For instance, Boriah et al. conducted a comparison study on similarity measures for categorical data and evaluated similarity measures in the context of outlier detection for categorical data [20]. It was concluded that the performance of an outlier detection algorithm is significantly affected by the similarity measure. In their research, it was not possible to introduce a best performing similarity measure, but they analyzed and reported the situations in which a measure has poor or superior performance. In another research work, Fernando et al. [21] reviewed, compared and benchmarked binary-based similarity measures for categorical data. With some cases studies, Deshpande et al. focused on data from a single knowledge area, for example biological data, and conducted a comparison in favor of profile similarity measures for genetic interaction networks. They concluded that the Dot Product is consistent among the best measures in different conditions and genetic interaction datasets [22].

Similarly, in the context of clustering, studies have been done on the effects of similarity measures., In one study Strehl and colleagues tried to recognize the impact of similarity measures on web clustering [23]. In another, six similarity measure were assessed, this time for trajectory clustering in outdoor surveillance scenes [24]. In chemical databases, Al Khalifa et. al. [25] examined performance of twelve coefficients for clustering, similarity searching and compound selection. From the results they concluded that no single coefficient is appropriate for all methodologies.

Despite these studies, no empirical analysis and comparison is available for clustering continuous data to investigate their behavior in low and high dimensional datasets. At the other hand our datasets are coming from a variety of applications and domains and while they are limited with a specific domain. In this study, we gather known similarity/distance measures available for clustering continuous data, which will be examined using various clustering algorithms and against 15 publicly available datasets. It is not possible to introduce a perfect similarity measure for all kinds of datasets, but in this paper we will discover the reaction of similarity measures to low and high-dimensional datasets. The similarity measures with the best results in each category are also introduced.

Before presenting the similarity measures for clustering continuous data, a definition of a clustering problem should be given. Assuming that the number of clusters required to be created is an input value k, the clustering problem is defined as follows [26]:

Definition 1

Given a dataset $D = \{v_1, v_2, \ldots, v_n\}$ of data vectors and an integer value k, the clustering problem is to define a mapping $f: D \rightarrow \{1, \ldots, k\}$ where each v_i is assigned to one cluster C_j, $1 \leq j \leq k$. A cluster C_j contains precisely those data vectors mapped to it; that is, $C_j = \{v_i \mid f(t_i) = C_j, 1 \leq i \leq n, \textit{and } v_i \in D\}$.

In the rest of this study, v_1, v_2 represent two data vectors defined as $v_1 = \{x_1, x_2, \ldots, x_n\}$, $v_2 = \{y_1, y_2, \ldots, y_n\}$, where x_i, y_i are called attributes.

Subsequently, similarity measures for clustering continuous data are discussed. Some of these similarity measures are frequently employed for clustering purposes while others have scarcely appeared in literature.

Minkowski

The Minkowski family includes Euclidean distance and Manhattan distance, which are particular cases of the Minkowski distance [27–29]. The Minkowski distance is defined by $d_{min} = (\sum_{i=1}^{n} |x_i - y_i|^m)^{\frac{1}{m}},\ m \geq 1$, where m is a positive real number and x_i and y_i are two vectors in n-dimensional space. The Minkowski distance performs well when the dataset clusters are isolated or compacted; if the dataset does not fulfil this condition, then the large-scale attributes would dominate the others [30,31]. Another problem with Minkowski metrics is that the largest-scale feature dominates the rest. Thus, normalizing the continuous features is the solution to this problem [31].

A modified version of the Minkowski metric has been proposed to solve clustering obstacles. For example, Wilson and Martinez presented distance based on counts for nominal attributes and a modified Minkowski metric for continuous features [32].

Manhattan distance

Manhattan distance is a special case of the Minkowski distance at m = 1. Like its parent, Manhattan is sensitive to outliers. When this distance measure is used in clustering algorithms, the shape of clusters is hyper-rectangular [33]. A study by Perlibakas demonstrated that a modified version of this distance measure is among the best distance measures for PCA-based face recognition [34]. This measure is defined as $d_{man} = \sum_{i=1}^{n} |x_i - y_i|$.

Euclidean distance

The most well-known distance used for numerical data is probably the Euclidean distance. This is a special case of the Minkowski distance when m = 2. Euclidean distance performs well when deployed to datasets that include compact or isolated clusters [30,31]. Although Euclidean distance is very common in clustering, it has a drawback: if two data vectors have no attribute values in common, they may have a smaller distance than the other pair of data vectors containing the same attribute values [31,35,36]. Another problem with Euclidean distance as a family of the Minkowski metric is that the largest-scaled feature would dominate the others. Normalization of continuous features is a solution to this problem [31].

Average distance

Regarding the above-mentioned drawback of Euclidean distance, average distance is a modified version of the Euclidean distance to improve the results [27,35]. For two data points x, y in *n*-dimentional space, the average distance is defined as $d_{ave} = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2\right)^{\frac{1}{2}}$.

Weighted euclidean distance

If the relative importance according to each attribute is available, then the Weighted Euclidean distance—another modification of Euclidean distance—can be used [37]. This distance is defined as $d_{we} = \left(\sum_{i=1}^{n} w_i(x_i - y_i)^2\right)^{\frac{1}{2}}$, where w_i is the weight given to the *i*th component.

This distance measure is the only measure which is not included in this study for comparison since calculating the weights is closely related to the dataset and the aim of researcher for cluster analysis on the dataset. As an instance of using this measure reader can refer to Ji et. al. research work. They used this measure for proposing a dynamic fuzzy cluster algorithm for time series [38].

Chord distance

Chord distance is one more Euclidean distance modification to overcome the previously mentioned Euclidean distance shortcomings. It can solve problems caused by the scale of measurements as well. Chord distance is defined as the length of the chord joining two normalized points within a hypersphere of radius one. This distance can be calculated from non-normalized data as well [27]. Chord distance is defined as $d_{chord} = \left(2 - 2\frac{\sum_{i=1}^{n} x_i y_i}{\|x\|_2 \|y\|_2}\right)^{\frac{1}{2}}$, where $\|x\|_2$ is the L^2-norm $\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$.

Mahalanobis distance

Mahalanobis distance is a data-driven measure in contrast to Euclidean and Manhattan distances that are independent of the related dataset to which two data points belong [20,33]. A regularized Mahalanobis distance can be used for extracting hyperellipsoidal clusters [30]. On the other hand, Mahalanobis distance can alleviated distortion caused by linear correlation among features by applying a whitening transformation to the data or by using the squared Mahalanobis distance [31]. Mahalanobis distance is defined by $d_{mah} = \sqrt{(x - y)S^{-1}(x - y)^T}$ where S is the covariance matrix of the dataset [27,39].

Cosine deasure

The Cosine similarity measure is mostly used in document similarity [28,33] and is defined as $Cosine(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{\|x\|_2 \|y\|_2}$, where $\|y\|_2$ is the Euclidean norm of vector $y = (y_1, y_2, \ldots, y_n)$ defined as $\|y\|_2 = \sqrt{y_1^2 + y_2^2 + \ldots + y_n^2}$. The Cosine measure is invariant to rotation but is variant to linear transformations. It is also independent of vector length [33].

Pearson correlation

Pearson correlation is widely used in clustering gene expression data [33,36,40]. This similarity measure calculates the similarity between the shapes of two gene expression patterns. The Pearson correlation is defined by $Pearson(x, y) = \frac{\sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}}$, where μ_x and μ_y are the means for x and y respectively. The Pearson correlation has a disadvantage of being sensitive to outliers [33,40].

The similarity measures explained above are the most commonly used for clustering continuous data. Table 1 represents a summary of these with some highlights of each.

Methodology of the Study

3.1 Experimental design

This section is devoted to explain the method and the framework which is used in this study for evaluating the effect of similarity measures on clustering quality. The main objective of this research study is to analyse the effect of different distance measures on quality of clustering algorithm results. As it is illustrated in Fig 1 there are 15 datasets used with 4 distance based algorithms on a total of 12 distance measures. All the distance measures in Table 1 are examined except the Weighted Euclidean distance which is dependent on the dataset and the aim of clustering.

Fig 2 explains the methodology of the study briefly. For each dataset we examined all four distance based algorithms, and each algorithms' quality of clustering has been evaluated by each 12 distance measures as it is demonstrated in Fig 1. It makes a total of 720 experiments in this research work to analyse the effect of distance measures. Representing and comparing this huge number of experiments is a challenging task and could not be done using ordinary charts and tables. Consequently we have developed a special illustration method using heat mapped tables in order to demonstrate all the results in the way that could be read and understand quickly. This method is described in section 4.1.1.

Table 1. Similarity Measures for continuous data (in time complexity, *n* is the number of dimensions of *x* and *y*).

Distance Measure	Equation	Time complexity	Advantages	Disadvantages	Applications
Euclidean Distance	$d_{euc} = \left[\sum_{i=1}^{n}(x_i - y_i)^2\right]^{\frac{1}{2}}$	O(n)	Very common, easy to compute and works well with datasets with compact or isolated clusters [27,31].	Sensitive to outliers [27,31].	*K*-means algorithm, Fuzzy *c*-means algorithm [38].
Average Distance	$d_{ave} = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2\right)^{\frac{1}{2}}$	O(n)	Better than Euclidean distance [35] at handling outliers.	Variables contribute independently to the measure of distance. Redundant values could dominate the similarity between data points [37].	*K*-means algorithm
Weighted Euclidean	$d_{we} = \left(\sum_{i=1}^{n}w_i(x_i - y_i)^2\right)^{\frac{1}{2}}$	O(n)	The weight matrix allows to increase the effect of more important data points than less important one [37].	Same as Average Distance.	Fuzzy *c*-means algorithm [38]
Chord	$d_{chord} = \left(2 - 2\frac{\sum_{i=1}^{n}x_i y_i}{\Vert x\Vert_2 \Vert y\Vert_2}\right)^{\frac{1}{2}}$	O(3n)	Can work with un-normalized data [27].	It is not invariant to linear transformation [33].	Ecological resemblance detection [35].
Mahalanobis	$\mathrm{d_{mah}} = \sqrt{(\mathrm{x} - \mathrm{y})\mathrm{S}^{-1}(\mathrm{x} - \mathrm{y})^{\mathrm{T}}}$	O(3n)	Mahalanobis is a data-driven measure that can ease the distance distortion caused by a linear combination of attributes [35].	It can be expensive in terms of computation [33]	Hyperellipsoidal clustering algorithm [30].
Cosine Measure	$\mathrm{Cosine}(\mathrm{x},\mathrm{y}) = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{x_i y_i}}{\Vert\mathrm{x}\Vert_2\Vert\mathrm{y}\Vert_2}$	O(3n)	Independent of vector length and invariant to rotation [33].	It is not invariant to linear transformation [33].	Mostly used in document similarity applications [28,33].
Manhattan	$d_{man} = \sum_{i=1}^{n}(x_i - y_i)$	O(n)	Is common and like other Minkowski-driven distances it works well with datasets with compact or isolated clusters [27].	Sensitive to the outliers. [27,31]	*K*-means algorithm
Mean Character Difference	$d_{MCD} = \frac{1}{n}\sum_{i=1}^{n}\lvert x_i - y_i\rvert$	O(n)	*Results in accurate outcomes using the K-medoids algorithm.	*Low accuracy for high-dimensional datasets using K-means.	Partitioning and hierarchical clustering algorithms.
Index of Association	$d_{IOA} = \frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{x_i}{\sum_{i=1}^{n}x_i} - \frac{y_i}{\sum_{i=1}^{n}y_i}\right\rvert$	O(3n)	-	*Low accuracy using K-means and K-medoids algorithms.	Partitioning and hierarchical clustering algorithms.
Canberra Metric	$d_{canb} = \sum_{i=1}^{n}\frac{\lvert x_i - y_i\rvert}{(x_i + y_i)}$	O(n)	*Results in accurate outcomes for high-dimensional datasets using the K-medoids algorithm.	-	Partitioning and hierarchical clustering algorithms.
Czekanowski Coefficient	$d_{czekan} = 1 - \frac{2\sum_{i=1}^{n}min(x_i, y_i)}{\sum_{i=1}^{n}(x_i + y_i)}$	O(2n)	*Results in accurate outcomes for medium-dimensional datasets using the K-means algorithm.	-	Partitioning and hierarchical clustering algorithms.
Coefficient of Divergence	$d_{canb} = \left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - y_i}{x_i + y_i}\right)^2\right)^{\frac{1}{2}}$	O(n)	*Results in accurate outcomes using the K-means algorithm.	-	Partitioning and hierarchical clustering algorithms.

(*Continued*)

Table 1. (*Continued*)

Distance Measure	Equation	Time complexity	Advantages	Disadvantages	Applications
Pearson coefficient	$Pearson(x,y) = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}}$	O(2n)	*Results in accurate outcomes using the hierarchical single-link algorithm for high dimensional datasets.	-	Partitioning and hierarchical clustering algorithms.

*Points marked by asterisk are compiled based on this article's experimental results.

doi:10.1371/journal.pone.0144059.t001

3.2 Rand Index

In this study, we used Rand Index (RI) for evaluation of clustering outcomes resulted by various distance measures. This section is an overview on this measure and it investigates the reason that this measure has been chosen.

Rand index is frequently used in measuring clustering quality. It is a measure of agreement between two sets of objects: first is the set produced by clustering process and the other defined by external criteria. Although there are different clustering measures such as Sum of Squared Error, Entropy, Purity, Jaccard etc. but among them the Rand index is probably the most used index for cluster validation [17,41,42]. Assuming $S = \{o_1, o_2, \ldots, o_n\}$ is a set of n elements and two partitions of S are given to compare $C = \{c_1, c_2, \ldots, c_r\}$, which is a partition of S into r subsets and $G = \{g_1, g_2, \ldots, g_s\}$, a partition of S into s subsets, the Rand index (R) is defined as follows:

Definition 2

$$RI = \frac{a+b}{a+b+c+d} \tag{1}$$

where:

- a is the number of pairs of vectors in S that are in the same set in C and in the same set in G.
- b is the number of pairs of elements in S that are in different sets in C and in different sets in G.
- c is the number of pairs of elements in S that are in the same set in C and in different sets in G.
- d is the number of pairs of elements in S that are in different sets in C and in the same set in G.

There is a modified version of rand index called Adjusted Rand Index (ARI) which is proposed by Hubert and Arabie [42] as an improvement for known problems with RI. These problems happen when the expected value of the RI of two random partition does not take a constant value (zero for example) or the Rand statistic approaches its upper limit of unity as the number of cluster increases. However, since our datasets don't have these problems and also owing to the fact that the results generated using ARI were following the same pattern of RI results, we have used Rand Index in this study due to its popularity in clustering community for clustering validation.

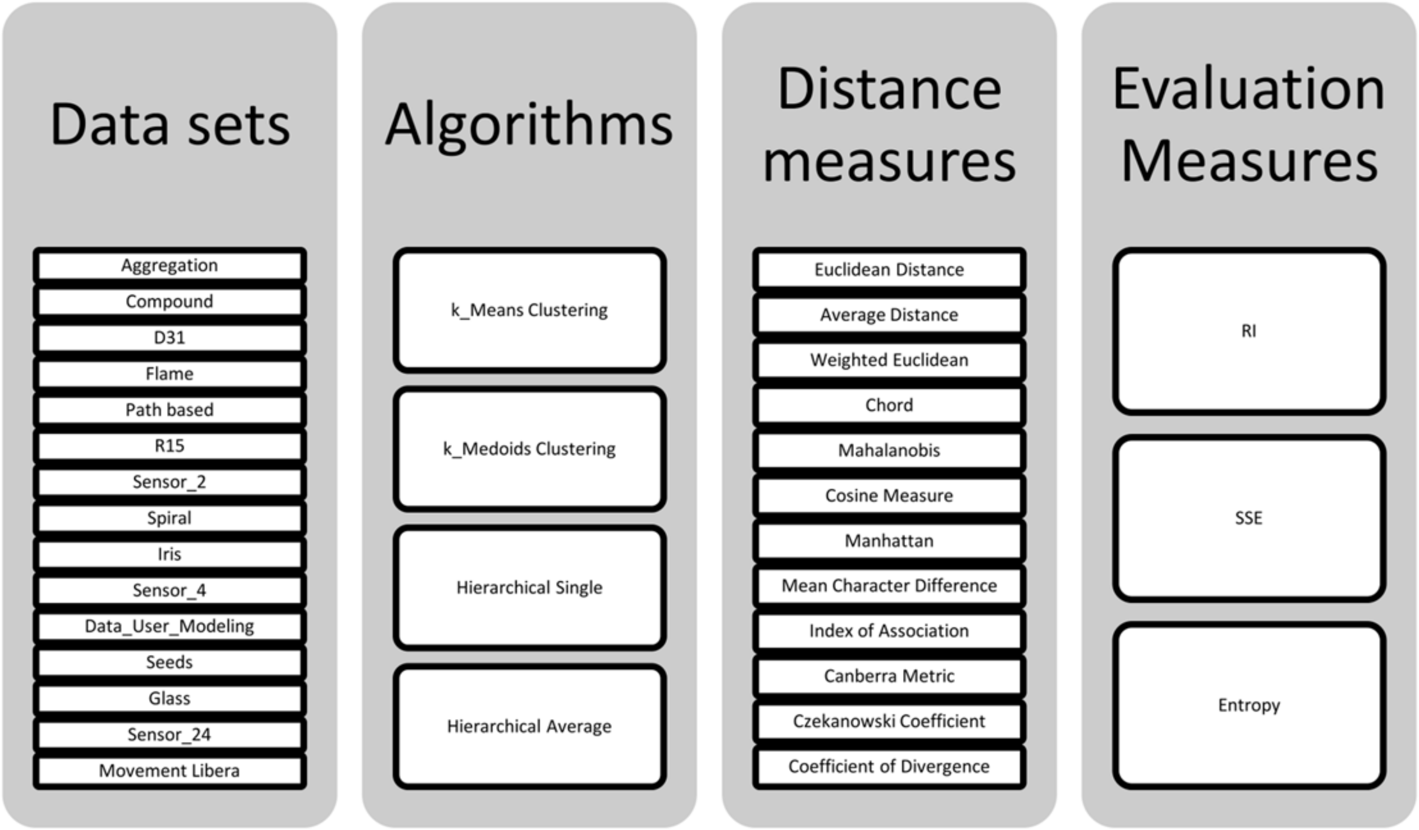


Fig 1. Overview of experimental study.

doi:10.1371/journal.pone.0144059.g001

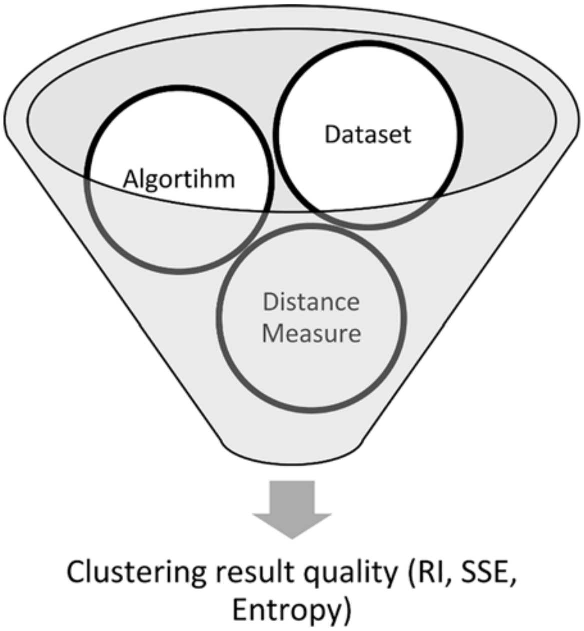


Fig 2. Arrangement of experiments.

doi:10.1371/journal.pone.0144059.g002

In this study we normalized the Rand Index values for the experiments. The normalized values are between 0 and 1 and we used following formula to approach it:

$$z_i = \frac{r_i - \min(r)}{\max(r) - \min(r)} \qquad 2$$

where $r = (r_1, \ldots, r_n)$ is the array of rand indexes produced by each similarity measure.

3.3 Analysis of variance (ANOVA) test

Before continuing this study, the main hypothesis needs to be proved: "distance measure has a considerable influence on clustering results". In order to show that distance measures cause significant difference on clustering quality, we have used ANOVA test. For this purpose we will consider a null hypothesis: "distance measures doesn't have significant influence on clustering quality". Using ANOVA test, if the p value be very small, it means that there is very small opportunity that null hypothesis is correct, and consequently we can reject it.

ANOVA analyzes the differences among a group of variable which is developed by Ronald Fisher [43]. ANOVA is a statistical test that demonstrate whether the mean of several groups are equal or not and it can be said that it generalizes the t-test for more than two groups. It is useful for testing means of more than two groups or variable for statistical significance. Statistical significance in statistics is achieved when a p-value is less than the significance level [44]. The p-value is the probability of obtaining results which acknowledge that the null hypothesis is true [45].

For ANOVA test we have considered a table with the structure shown in Table 2 which covers all RI results for all four algorithms and each distance/similarity measure and for all datasets. Table is divided into 4 section for four respective algorithms. In each sections rows represent results generated with distance measures for a dataset.

ANOVA test is performed for each algorithm separately to find if distance measures have significant impact on clustering results in each clustering algorithm.

The ANOVA test result on above table is demonstrated in the Tables 3–6.

The small Prob values indicates that differences between means of the columns are significant. From that we can conclude that the similarity measures have significant impact in clustering quality. In the rest of this study we will inspect how these similarity measures influence on clustering quality.

Experimental Results

It is noted that references to all data employed in this work are available in acknowledgment section. A diverse set of similarity measures for continuous data was studied on low and high-dimensional continuous datasets in order to clarify and compare the accuracy of each similarity measure in different datasets with various dimensionality situations and using 15 datasets [18,19,46–49]. Details of the datasets applied in this study are represented in Table 7.

The experiments were conducted using partitioning (k-means and k-medoids) and hierarchical algorithms, which are distance-based. As it is discussed in section 3.2 the Rand index served to evaluate and compare the results. The results for each of these algorithms are discussed later in this section.

The k-means and k-medoids algorithms were used in this experiment as partitioning algorithms, and the Rand index served accuracy evaluation purposes. Due to the fact that the k-means and k-medoids algorithm results are dependent on the initial, randomly selected centers, and in some cases their accuracy might be affected by local minimum trap, the experiment

Table 2. Rand Index values used for ANOVA test (in the table HAverage stands for Hierarchical Average algorithm and HSingle stands for Hierarchical Single link).

Dataset	Distance/Similarity Measures											
	Euclidean	Average	Cosine	Chord	Mahalanobis	Canberra	CoeffDiv	Czekan	IndOfAssoc	Manhattan	MCharDiff	Pearson
						k-Means						
sensor_2	0.722	0.733	0.659	0.659	0.725	0.744	0.741	0.765	0.662	0.729	0.729	0.403
Aggregation	0.929	0.929	0.798	0.799	0.927	0.921	0.904	0.949	0.799	0.927	0.927	0.636
Compound	0.919	0.914	0.746	0.746	0.926	0.890	0.908	0.886	0.744	0.906	0.904	0.497
Flame	0.756	0.756	0.569	0.569	0.750	0.716	0.498	0.710	0.557	0.750	0.750	0.536
Pathbased	0.750	0.750	0.639	0.639	0.758	0.735	0.733	0.746	0.637	0.748	0.748	0.635
R15	0.999	0.999	0.949	0.948	0.999	0.999	0.998	0.998	0.947	0.998	0.998	0.552
Spiral	0.554	0.554	0.562	0.562	0.555	0.550	0.552	0.553	0.562	0.556	0.556	0.496
D31	0.994	0.992	0.956	0.956	0.995	0.992	0.992	0.994	0.956	0.994	0.994	0.528
Iris	0.880	0.880	0.966	0.966	0.880	0.942	0.950	0.927	0.958	0.874	0.874	0.776
sensor_4	0.612	0.624	0.637	0.637	0.619	0.745	0.709	0.737	0.649	0.726	0.728	0.670
Data_User_Modeling	0.725	0.725	0.668	0.668	0.719	0.711	0.706	0.713	0.668	0.712	0.711	0.657
Seeds	0.876	0.874	0.884	0.884	0.876	0.859	0.782	0.891	0.890	0.872	0.872	0.359
Glass	0.741	0.742	0.737	0.740	0.732	0.604	0.602	0.734	0.732	0.734	0.731	0.342
sensor_24	0.610	0.615	0.614	0.617	0.596	0.618	0.621	0.613	0.610	0.604	0.611	0.626
Libras movement	0.914	0.917	0.913	0.917	0.915	0.911	0.914	0.910	0.913	0.914	0.912	0.918
						k-Medoids						
sensor_2	0.777	0.736	0.661	0.661	0.729	0.804	0.806	0.797	0.675	0.785	0.796	0.403
Aggregation	0.949	0.949	0.790	0.790	0.950	0.928	0.901	0.958	0.787	0.941	0.953	0.636
Compound	0.925	0.911	0.734	0.733	0.920	0.890	0.890	0.900	0.740	0.916	0.913	0.497
Flame	0.762	0.762	0.538	0.538	0.756	0.705	0.498	0.716	0.565	0.744	0.744	0.536
Pathbased	0.746	0.746	0.606	0.606	0.756	0.743	0.745	0.745	0.667	0.741	0.741	0.635
R15	0.999	0.999	0.947	0.945	0.988	0.998	0.988	0.998	0.947	0.999	0.998	0.552
Spiral	0.555	0.554	0.555	0.555	0.555	0.571	0.555	0.557	0.551	0.556	0.564	0.496
D31	0.994	0.992	0.956	0.956	0.992	0.990	0.988	0.991	0.956	0.991	0.994	0.528
Iris	0.912	0.912	0.966	0.966	0.824	0.927	0.950	0.906	0.950	0.880	0.880	0.776
sensor_4	0.707	0.711	0.711	0.711	0.656	0.740	0.722	0.709	0.690	0.696	0.716	0.656
Data_User_Modeling	0.725	0.712	0.654	0.654	0.728	0.285	0.285	0.285	0.646	0.734	0.745	0.659
Seeds	0.874	0.874	0.842	0.842	0.798	0.872	0.771	0.876	0.865	0.867	0.867	0.359
Glass	0.735	0.736	0.738	0.732	0.711	0.633	0.582	0.737	0.735	0.737	0.739	0.342
sensor_24	0.624	0.623	0.623	0.622	0.588	0.652	0.634	0.630	0.629	0.620	0.617	0.613
Libras movement	0.907	0.909	0.908	0.905	0.720	0.897	0.905	0.901	0.906	0.904	0.904	0.907
						HSingle						
sensor_2	0.432	0.432	0.355	0.355	0.432	0.432	0.432	0.431	0.365	0.432	0.432	0.405
Aggregation	0.926	0.926	0.574	0.574	0.926	0.619	0.927	0.927	0.550	0.926	0.926	0.635
Compound	0.890	0.890	0.415	0.415	0.896	0.895	0.898	0.891	0.415	0.712	0.712	0.497
Flame	0.541	0.541	0.522	0.522	0.541	0.531	0.531	0.541	0.522	0.541	0.541	0.538
Pathbased	0.338	0.338	0.362	0.362	0.340	0.339	0.338	0.338	0.362	0.338	0.338	0.635
R15	0.910	0.910	0.817	0.817	0.910	0.856	0.857	0.856	0.817	0.911	0.911	0.574
Spiral	1.000	1.000	0.383	0.383	1.000	0.781	0.781	0.781	0.383	1.000	1.000	0.497

(*Continued*)

Table 2. (*Continued*)

Dataset	Distance/Similarity Measures											
	Euclidean	Average	Cosine	Chord	Mahalanobis	Canberra	CoeffDiv	Czekan	IndOfAssoc	Manhattan	MCharDiff	Pearson
D31	0.779	0.779	0.818	0.818	0.754	0.740	0.731	0.730	0.518	0.755	0.755	0.536
Iris	0.777	0.777	0.772	0.772	0.343	0.753	0.753	0.772	0.772	0.776	0.776	0.772
sensor_4	0.341	0.341	0.345	0.345	0.346	0.451	0.339	0.333	0.345	0.338	0.338	0.651
Data_User_Modeling	0.309	0.309	0.301	0.301	0.304	0.302	0.302	0.305	0.302	0.299	0.299	0.311
Seeds	0.357	0.357	0.340	0.340	0.337	0.340	0.337	0.340	0.340	0.340	0.340	0.358
Glass	0.304	0.304	0.308	0.308	0.309	0.293	0.294	0.308	0.308	0.308	0.308	0.342
sensor_24	0.347	0.347	0.346	0.346	0.353	0.346	0.347	0.346	0.346	0.345	0.345	0.349
Libras movement	0.187	0.187	0.202	0.202	0.131	0.183	0.183	0.187	0.192	0.187	0.187	0.296
HAverage												
sensor_2	0.466	0.466	0.634	0.634	0.506	0.466	0.729	0.716	0.634	0.466	0.466	0.404
Aggregation	1.000	1.000	0.778	0.778	0.997	0.930	0.948	0.927	0.778	0.991	0.991	0.643
Compound	0.921	0.921	0.676	0.676	0.921	0.850	0.852	0.829	0.697	0.933	0.933	0.511
Flame	0.721	0.721	0.503	0.503	0.847	0.512	0.529	0.501	0.503	0.689	0.689	0.538
Pathbased	0.738	0.738	0.699	0.699	0.754	0.438	0.377	0.708	0.629	0.724	0.724	0.635
R15	0.999	0.999	0.917	0.917	0.999	0.981	0.963	0.990	0.914	0.998	0.998	0.566
Spiral	0.537	0.537	0.528	0.528	0.557	0.424	0.499	0.498	0.428	0.540	0.540	0.497
D31	0.994	0.994	0.950	0.950	0.996	0.977	0.979	0.986	0.952	0.996	0.996	0.537
Iris	0.892	0.892	0.772	0.772	0.343	0.753	0.753	0.778	0.772	0.886	0.886	0.776
sensor_4	0.338	0.338	0.561	0.561	0.338	0.479	0.479	0.480	0.544	0.376	0.376	0.653
Data_User_Modeling	0.659	0.659	0.301	0.301	0.337	0.302	0.302	0.307	0.309	0.645	0.645	0.594
Seeds	0.887	0.887	0.691	0.691	0.337	0.879	0.581	0.802	0.688	0.802	0.802	0.362
Glass	0.329	0.329	0.570	0.570	0.309	0.328	0.323	0.415	0.415	0.415	0.415	0.369
sensor_24	0.353	0.353	0.538	0.538	0.347	0.498	0.516	0.518	0.521	0.428	0.428	0.446
Libras movement	0.886	0.886	0.892	0.892	0.131	0.582	0.613	0.827	0.844	0.861	0.861	0.886

doi:10.1371/journal.pone.0144059.t002

Table 3. ANOVA results for k-means.

K_means	SS	df	MS	F	Prob>F
Columns	0.68317	11	0.06211	2.96	0.0013
Error	3.52624	168	0.02099		
Total	4.20942	179			

doi:10.1371/journal.pone.0144059.t003

was repeated 100 times for each similarity measure, after which the maximum Rand index was considered for comparison.

4.1 Illustration technique

A summary of the normalized Rand index results is illustrated in color scale tables in Fig 3 and Fig 4. Since the aim of this study is to investigate and evaluate the accuracy of similarity measures for different dimensional datasets, the tables are organized based on horizontally ascending dataset dimensions. After the first column, which contains the names of the similarity measures, the remaining table is divided in two batches of columns (low and high-dimensional) that demonstrate the normalized Rand indexes for low and high-dimensional datasets, respectively. The final column considered in this table is 'overall average' in order to explore the most accurate similarity measure in general. This illustrational structure and approach is used for all four algorithms in this paper.

4.2 Benchmarking similarity measures for partitioning algorithms

Fig 3 represents the results for the k-means algorithm. According to the figure, for low-dimensional datasets, the Mahalanobis measure has the highest results among all similarity measures. On the other hand, for high-dimensional datasets, the Coefficient of Divergence is the most accurate with the highest Rand index values. Fig 4 provides the results for the k-medoids algorithm. Mean Character Difference is the most precise measure for low-dimensional datasets, while the Cosine measure represents better results in terms of accuracy for high-dimensional datasets. Overall, Mean Character Difference has high accuracy for most datasets.

As a general result for the partitioning algorithms used in this study, average distance results in more accurate and reliable outcomes for both algorithms. It is the most accurate measure in the k-means algorithm and at the same time, with very little difference, it stands in second place after Mean Character Difference for the k-medoids algorithm.

From another perspective, similarity measures in the k-means algorithm can be investigated to clarify which would lead to the k-means converging faster. However the convergence of k-means and k-medoid algorithms is not guaranteed due to the possibility of falling in local minimum trap. For this reason we have run the algorithm 100 times to prevent bias toward this weakness. Fig 5 shows two sample box charts created by using normalized data, which represents the normalized iteration count needed for the convergence of each similarity measure.

Table 4. ANOVA results for k-medoids.

K_medoids	SS	df	MS	F	Prob>F
Columns	0.69565	11	0.06324	2.62	0.0042
Error	4.05766	168	0.02415		
Total	4.75331	179			

doi:10.1371/journal.pone.0144059.t004

Table 5. ANOVA results for HSingle.

HAvrage	SS	df	MS	F	Prob>F
Columns	0.47251	11	0.04296	2.62	0.0043
Error	2.52617	154	0.0164		
Total	8.91175	175			

doi:10.1371/journal.pone.0144059.t005

Table 6. ANOVA results for HSingle.

HSingle	SS	df	MS	F	Prob>F
Columns	0.3194	11	0.02903	2.38	0.0095
Error	1.8788	154	0.0122		
Total	10.2233	179			

doi:10.1371/journal.pone.0144059.t006

Results were collected after 100 times of repeating the k-means algorithm for each similarity measure and dataset.

Fig 6 is a summarized color scale table representing the mean and variance of iteration counts for all 100 algorithm runs. Pearson has the fastest convergence in most datasets. After Pearson, Average is the fastest similarity measure in terms of convergence.

Regarding the discussion on Rand index and iteration count, it is manifested that the Average measure is not only accurate in most datasets and with both k-means and k-medoids algorithms, but it is the second fastest similarity measure after Pearson in terms of convergence, making it a secure choice when clustering is necessary using k-means or k-medoids algorithms.

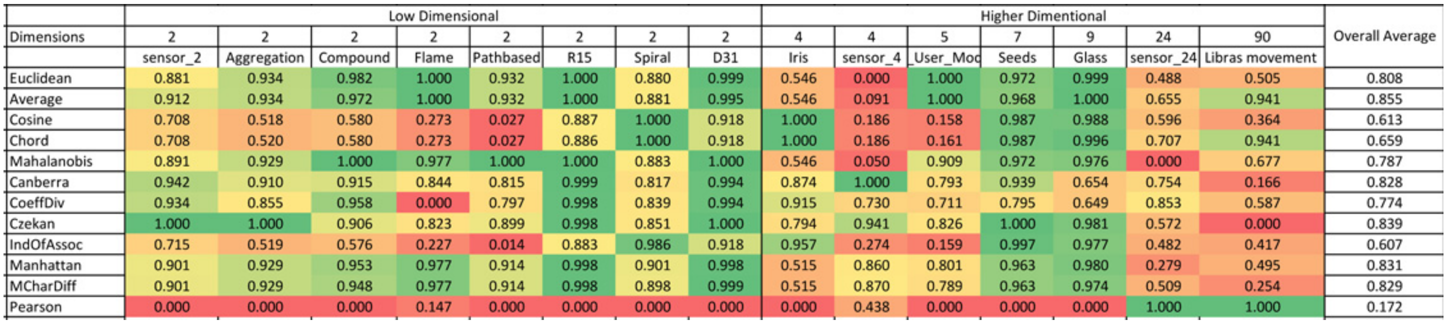
4.3 Benchmarking similarity measures for hierarchical algorithms

In a previous section, the influence of different similarity measures on k-means and k-medoids algorithms as partitioning algorithms was evaluated and compared. In this section, the results for Single-link and Group Average algorithms, which are two hierarchical clustering

Table 7. Dataset Details.

Dataset Name	Dimensions	Clusters	Vectors
Aggregation	2	7	788
Compound	2	6	399
D31	2	31	3100
Flame	2	2	240
Path based	2	3	300
R15	2	15	600
Sensor_2	2	4	5456
Spiral	2	3	312
Iris	4	3	150
Sensor_4	4	4	5456
Data_User_Modeling	5	4	258
Seeds	7	3	210
Glass	9	7	214
Sensor_24	24	4	5456
Movement Libera	90	15	360

doi:10.1371/journal.pone.0144059.t007



	Low Dimensional								Higher Dimentional							Overall Average
Dimensions	2	2	2	2	2	2	2	2	4	4	5	7	9	24	90	
	sensor_2	Aggregation	Compound	Flame	Pathbased	R15	Spiral	D31	Iris	sensor_4	_User_Moc	Seeds	Glass	sensor_24	Libras movement	
Euclidean	0.881	0.934	0.982	1.000	0.932	1.000	0.880	0.999	0.546	0.000	1.000	0.972	0.999	0.488	0.505	0.808
Average	0.912	0.934	0.972	1.000	0.932	1.000	0.881	0.995	0.546	0.091	1.000	0.968	1.000	0.655	0.941	0.855
Cosine	0.708	0.518	0.580	0.273	0.027	0.887	1.000	0.918	1.000	0.186	0.158	0.987	0.988	0.596	0.364	0.613
Chord	0.708	0.520	0.580	0.273	0.027	0.886	1.000	0.918	1.000	0.186	0.161	0.987	0.996	0.707	0.941	0.659
Mahalanobis	0.891	0.929	1.000	0.977	1.000	1.000	0.883	1.000	0.546	0.050	0.909	0.972	0.976	0.000	0.677	0.787
Canberra	0.942	0.910	0.915	0.844	0.815	0.999	0.817	0.994	0.874	1.000	0.793	0.939	0.654	0.754	0.166	0.828
CoeffDiv	0.934	0.855	0.958	0.000	0.797	0.998	0.839	0.994	0.915	0.730	0.711	0.795	0.649	0.853	0.587	0.774
Czekan	1.000	1.000	0.906	0.823	0.899	0.998	0.851	1.000	0.794	0.941	0.826	1.000	0.981	0.572	0.000	0.839
IndOfAssoc	0.715	0.519	0.576	0.227	0.014	0.883	0.986	0.918	0.957	0.274	0.159	0.997	0.977	0.482	0.417	0.607
Manhattan	0.901	0.929	0.953	0.977	0.914	0.998	0.901	0.998	0.515	0.860	0.801	0.963	0.980	0.279	0.495	0.831
MCharDiff	0.901	0.929	0.948	0.977	0.914	0.998	0.898	0.999	0.515	0.870	0.789	0.963	0.974	0.509	0.254	0.829
Pearson	0.000	0.000	0.000	0.147	0.000	0.000	0.000	0.000	0.000	0.438	0.000	0.000	0.000	1.000	1.000	0.172

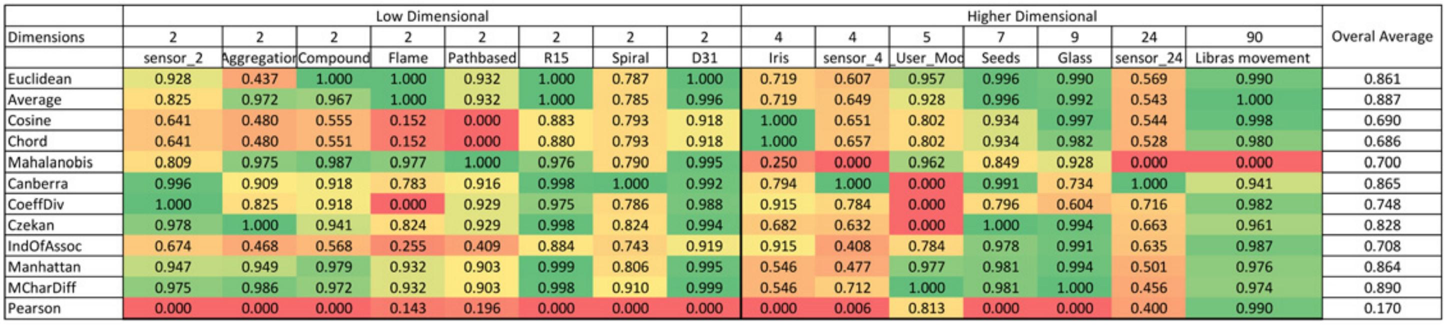
Fig 3. K-means color scale table for normalized Rand index values (green represents the highest and it changes to red, which is the lowest Rand index value).

doi:10.1371/journal.pone.0144059.g003

algorithms, will be discussed for each similarity measure in terms of the Rand index. Fig 7 and Fig 8 represent sample bar charts of the results. The bar charts include 6 sample datasets. Because bar charts for all datasets and similarity measures would be jumbled, the results are presented using color scale tables for easier understanding and discussion. As discussed in the last section, Fig 9 and Fig 10 are two color scale tables that demonstrate the normalized Rand index values for each similarity measure. The results in Fig 9 for Single-link show that for low-dimensional datasets, the Mahalanobis distance is the most accurate similarity measure and Pearson is the best among other measures for high-dimensional datasets. The overall average column in this figure shows that generally, Pearson presents the highest accuracy and the Average and Euclidean distances are among the most accurate measures. For the Group Average algorithm, as seen in Fig 10, Euclidean and Average are the best among all similarity measures for low-dimensional datasets. For high-dimensional datasets, Cosine and Chord are the most accurate measures. Generally, in the Group Average algorithm, Manhattan and Mean Character Difference have the best overall Rand index results followed by Euclidean and Average. Considering the overall results, it is clear that the Average measure is constantly among the best measures, and for both Single-link and Group Average algorithms.

A review of the results and discussions on the k-means, k-medoids, Single-link and Group Average algorithms reveals that by considering the overall results, the Average measure is regularly among the most accurate measures for all four algorithms.

According to heat map tables it is noticeable that Pearson correlation is behaving differently in comparison to other distance measures. It specially shows very weak results with centroid based algorithms, k-means and k-medoids. Based on the results in this research, in general,



	Low Dimensional								Higher Dimensional							Overal Average
Dimensions	2	2	2	2	2	2	2	2	4	4	5	7	9	24	90	
	sensor_2	Aggregatior	Compound	Flame	Pathbased	R15	Spiral	D31	Iris	sensor_4	_User_Moc	Seeds	Glass	sensor_24	Libras movement	
Euclidean	0.928	0.437	1.000	1.000	0.932	1.000	0.787	1.000	0.719	0.607	0.957	0.996	0.990	0.569	0.990	0.861
Average	0.825	0.972	0.967	1.000	0.932	1.000	0.785	0.996	0.719	0.649	0.928	0.996	0.992	0.543	1.000	0.887
Cosine	0.641	0.480	0.555	0.152	0.000	0.883	0.793	0.918	1.000	0.651	0.802	0.934	0.997	0.544	0.998	0.690
Chord	0.641	0.480	0.551	0.152	0.000	0.880	0.793	0.918	1.000	0.657	0.802	0.934	0.982	0.528	0.980	0.686
Mahalanobis	0.809	0.975	0.987	0.977	1.000	0.976	0.790	0.995	0.250	0.000	0.962	0.849	0.928	0.000	0.000	0.700
Canberra	0.996	0.909	0.918	0.783	0.916	0.998	1.000	0.992	0.794	1.000	0.000	0.991	0.734	1.000	0.941	0.865
CoeffDiv	1.000	0.825	0.918	0.000	0.929	0.975	0.786	0.988	0.915	0.784	0.000	0.796	0.604	0.716	0.982	0.748
Czekan	0.978	1.000	0.941	0.824	0.929	0.998	0.824	0.994	0.682	0.632	0.000	1.000	0.994	0.663	0.961	0.828
IndOfAssoc	0.674	0.468	0.568	0.255	0.409	0.884	0.743	0.919	0.915	0.408	0.784	0.978	0.991	0.635	0.987	0.708
Manhattan	0.947	0.949	0.979	0.932	0.903	0.999	0.806	0.995	0.546	0.477	0.977	0.981	0.994	0.501	0.976	0.864
MCharDiff	0.975	0.986	0.972	0.932	0.903	0.998	0.910	0.999	0.546	0.712	1.000	0.981	1.000	0.456	0.974	0.890
Pearson	0.000	0.000	0.000	0.143	0.196	0.000	0.000	0.000	0.000	0.006	0.813	0.000	0.000	0.400	0.990	0.170

Fig 4. K-medoids color scale table for normalized Rand index values (green is the highest and changes color to red, which is the lowest Rand index value).

doi:10.1371/journal.pone.0144059.g004

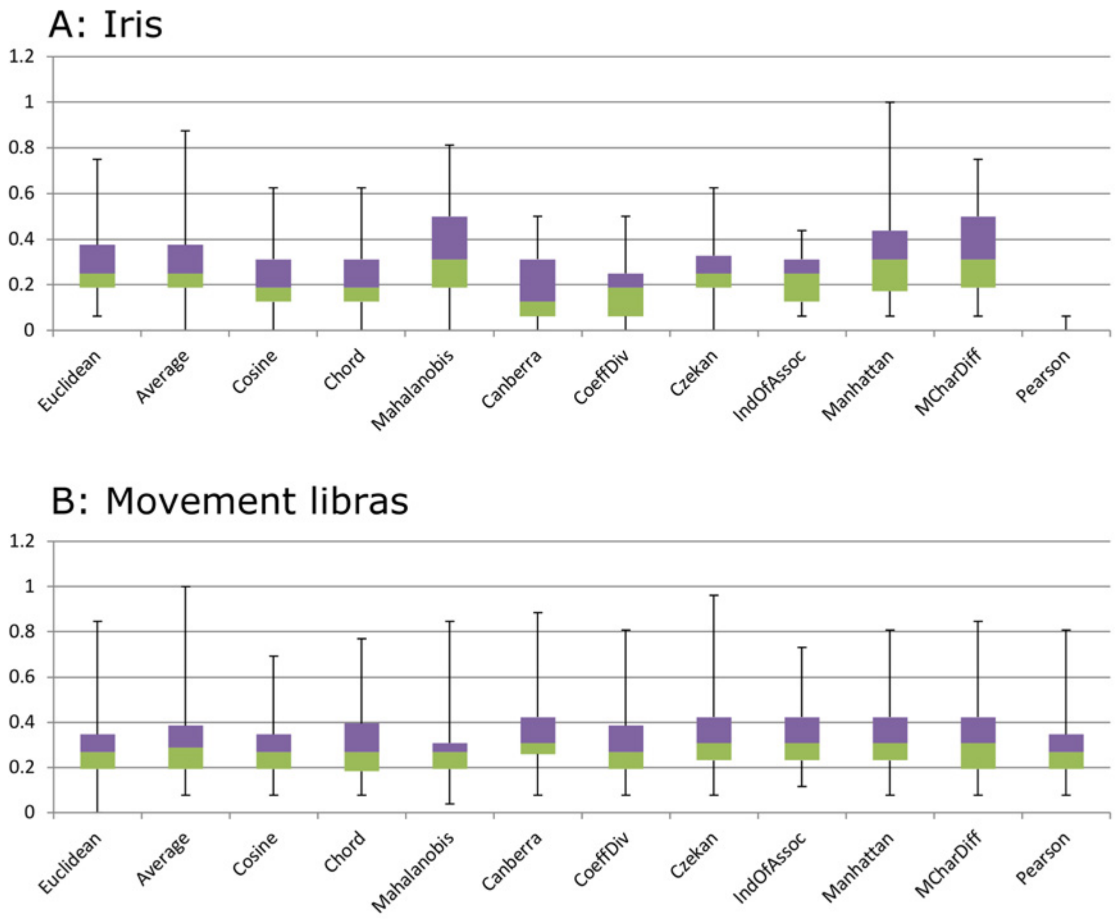


Fig 5. Sample box charts for k-means iteration counts created with a collection of normalized results after 100 times of repeating the algorithm for each similarity measure and dataset.

doi:10.1371/journal.pone.0144059.g005

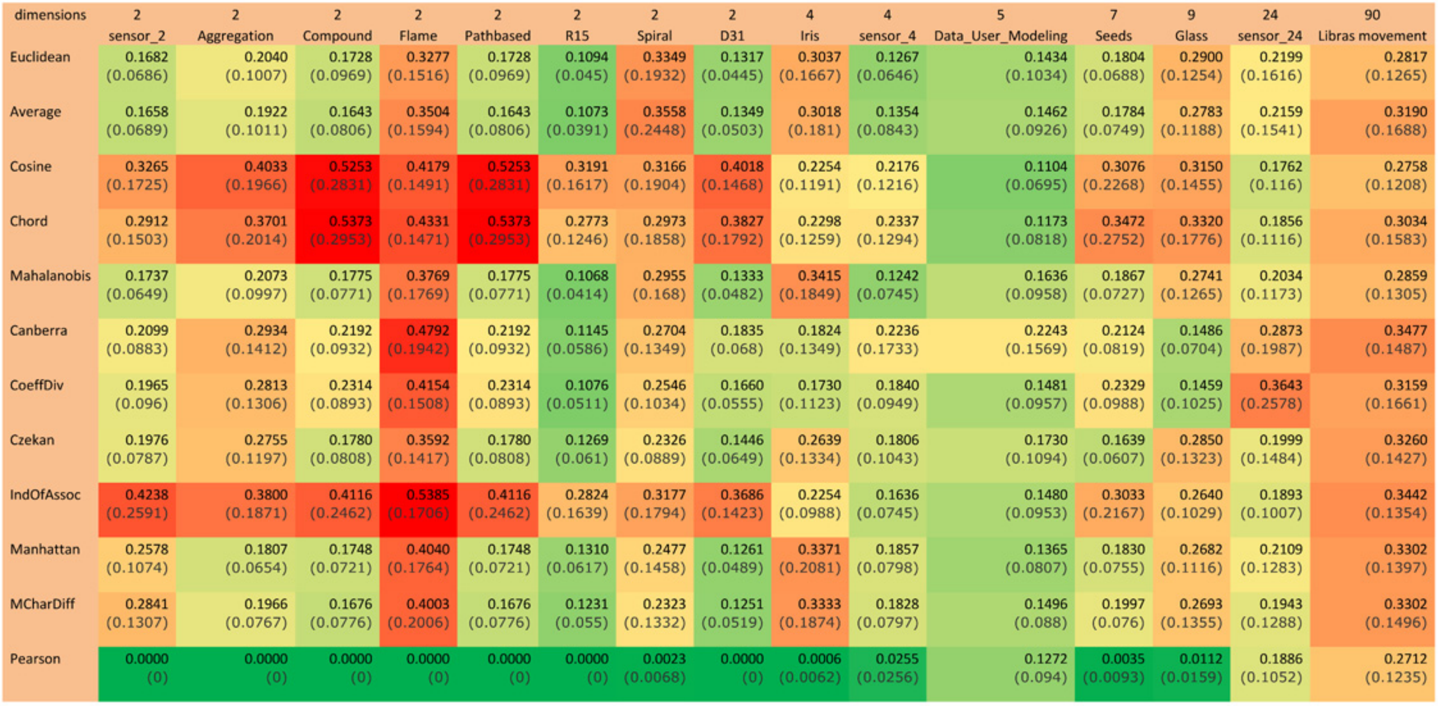
Pearson correlation doesn't work properly for low dimensional datasets while it shows better results for high dimensional datasets.

Fig 11 illustrates the overall average RI in all 4 algorithms and all 15 datasets also uphold the same conclusion. Fig 12 at the other hand shows the average RI for 4 algorithms separately. It can be inferred that Average measure among other measures is more accurate.

Furthermore, by using the k-means algorithm, this similarity measure is the fastest after Pearson in terms of convergence.

Concluding Remarks

Selecting the right distance measure is one of the challenges encountered by professionals and researchers when attempting to deploy a distance-based clustering algorithm to a dataset. The variety of similarity measures can cause confusion and difficulties in choosing a suitable measure. Similarity measures may perform differently for datasets with diverse dimensionalities. The aim of this study was to clarify which similarity measures are more appropriate for low-dimensional and which perform better for high-dimensional datasets in the experiments. In this work, similarity measures for clustering numerical data in distance-based algorithms were compared and benchmarked using 15 datasets categorized as low and high-dimensional



dimensions	2	2	2	2	2	2	2	2	4	4	5	7	9	24	90
	sensor_2	Aggregation	Compound	Flame	Pathbased	R15	Spiral	D31	Iris	sensor_4	Data_User_Modeling	Seeds	Glass	sensor_24	Libras movement
Euclidean	0.1682 (0.0686)	0.2040 (0.1007)	0.1728 (0.0969)	0.3277 (0.1516)	0.1728 (0.0969)	0.1094 (0.045)	0.3349 (0.1932)	0.1317 (0.0445)	0.3037 (0.1667)	0.1267 (0.0646)	0.1434 (0.1034)	0.1804 (0.0688)	0.2900 (0.1254)	0.2199 (0.1616)	0.2817 (0.1265)
Average	0.1658 (0.0689)	0.1922 (0.1011)	0.1643 (0.0806)	0.3504 (0.1594)	0.1643 (0.0806)	0.1073 (0.0391)	0.3558 (0.2448)	0.1349 (0.0503)	0.3018 (0.181)	0.1354 (0.0843)	0.1462 (0.0926)	0.1784 (0.0749)	0.2783 (0.1188)	0.2159 (0.1541)	0.3190 (0.1688)
Cosine	0.3265 (0.1725)	0.4033 (0.1966)	0.5253 (0.2831)	0.4179 (0.1491)	0.5253 (0.2831)	0.3191 (0.1617)	0.3166 (0.1904)	0.4018 (0.1468)	0.2254 (0.1191)	0.2176 (0.1216)	0.1104 (0.0695)	0.3076 (0.2268)	0.3150 (0.1455)	0.1762 (0.116)	0.2758 (0.1208)
Chord	0.2912 (0.1503)	0.3701 (0.2014)	0.5373 (0.2953)	0.4331 (0.1471)	0.5373 (0.2953)	0.2773 (0.1246)	0.2973 (0.1858)	0.3827 (0.1792)	0.2298 (0.1259)	0.2337 (0.1294)	0.1173 (0.0818)	0.3472 (0.2752)	0.3320 (0.1776)	0.1856 (0.1116)	0.3034 (0.1583)
Mahalanobis	0.1737 (0.0649)	0.2073 (0.0997)	0.1775 (0.0771)	0.3769 (0.1769)	0.1775 (0.0771)	0.1068 (0.0414)	0.2955 (0.168)	0.1333 (0.0482)	0.3415 (0.1849)	0.1242 (0.0745)	0.1636 (0.0958)	0.1867 (0.0727)	0.2741 (0.1265)	0.2034 (0.1173)	0.2859 (0.1305)
Canberra	0.2099 (0.0883)	0.2934 (0.1412)	0.2192 (0.0932)	0.4792 (0.1942)	0.2192 (0.0932)	0.1145 (0.0586)	0.2704 (0.1349)	0.1835 (0.068)	0.1824 (0.1349)	0.2236 (0.1733)	0.2243 (0.1569)	0.2124 (0.0819)	0.1486 (0.0704)	0.2873 (0.1987)	0.3477 (0.1487)
CoeffDiv	0.1965 (0.096)	0.2813 (0.1306)	0.2314 (0.0893)	0.4154 (0.1508)	0.2314 (0.0893)	0.1076 (0.0511)	0.2546 (0.1034)	0.1660 (0.0555)	0.1730 (0.1123)	0.1840 (0.0949)	0.1481 (0.0957)	0.2329 (0.0988)	0.1459 (0.1025)	0.3643 (0.2578)	0.3159 (0.1661)
Czekan	0.1976 (0.0787)	0.2755 (0.1197)	0.1780 (0.0808)	0.3592 (0.1417)	0.1780 (0.0808)	0.1269 (0.061)	0.2326 (0.0889)	0.1446 (0.0649)	0.2639 (0.1334)	0.1806 (0.1043)	0.1730 (0.1094)	0.1639 (0.0607)	0.2850 (0.1323)	0.1999 (0.1484)	0.3260 (0.1427)
IndOfAssoc	0.4238 (0.2591)	0.3800 (0.1871)	0.4116 (0.2462)	0.5385 (0.1706)	0.4116 (0.2462)	0.2824 (0.1639)	0.3177 (0.1794)	0.3686 (0.1423)	0.2254 (0.0988)	0.1636 (0.0745)	0.1480 (0.0953)	0.3033 (0.2167)	0.2640 (0.1029)	0.1893 (0.1007)	0.3442 (0.1354)
Manhattan	0.2578 (0.1074)	0.1807 (0.0654)	0.1748 (0.0721)	0.4040 (0.1764)	0.1748 (0.0721)	0.1310 (0.0617)	0.2477 (0.1458)	0.1261 (0.0489)	0.3371 (0.2081)	0.1857 (0.0798)	0.1365 (0.0807)	0.1830 (0.0755)	0.2682 (0.1116)	0.2109 (0.1283)	0.3302 (0.1397)
MCharDiff	0.2841 (0.1307)	0.1966 (0.0767)	0.1676 (0.0776)	0.4003 (0.2006)	0.1676 (0.0776)	0.1231 (0.055)	0.2323 (0.1332)	0.1251 (0.0519)	0.3333 (0.1874)	0.1828 (0.0797)	0.1496 (0.088)	0.1997 (0.076)	0.2693 (0.1355)	0.1943 (0.1288)	0.3302 (0.1496)
Pearson	0.0000 (0)	0.0000 (0)	0.0000 (0)	0.0000 (0)	0.0000 (0)	0.0000 (0)	0.0023 (0.0068)	0.0000 (0)	0.0006 (0.0062)	0.0255 (0.0256)	0.1272 (0.094)	0.0035 (0.0093)	0.0112 (0.0159)	0.1886 (0.1052)	0.2712 (0.1235)

Fig 6. Color scale table for iteration count mean and variance (green is the lowest and it changes color to red, which shows the greatest iteration count value).

doi:10.1371/journal.pone.0144059.g006

datasets. The accuracy of similarity measures in terms of the Rand index was studied and the best similarity measures for each of the low and high-dimensional datasets were discussed for four well-known distance-based algorithms. Overall, the results indicate that Average Distance is among the top most accurate measures for all clustering algorithms employed in this article. Moreover, this measure is one of the fastest in terms of convergence when k-means is the target clustering algorithm. Based on results in this study, in general, Pearson correlation is not recommended for low dimensional datasets. It also is not compatible with centroid based algorithms. However, this measure is mostly recommended for high dimensional datasets and by using hierarchical approaches.

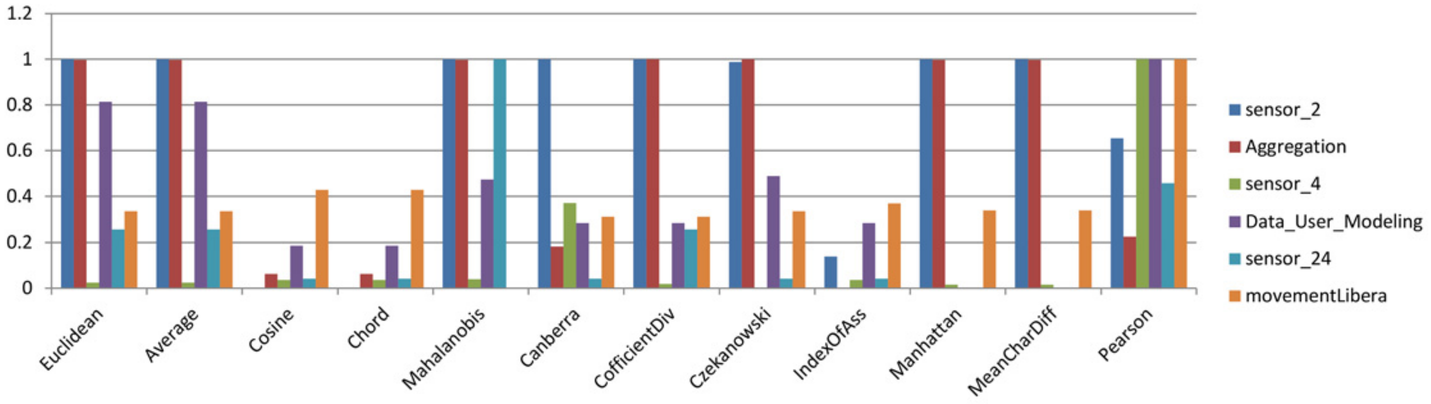


Fig 7. Bar chart of normalized Rand index values for selected datasets using the Single-link algorithm.

doi:10.1371/journal.pone.0144059.g007

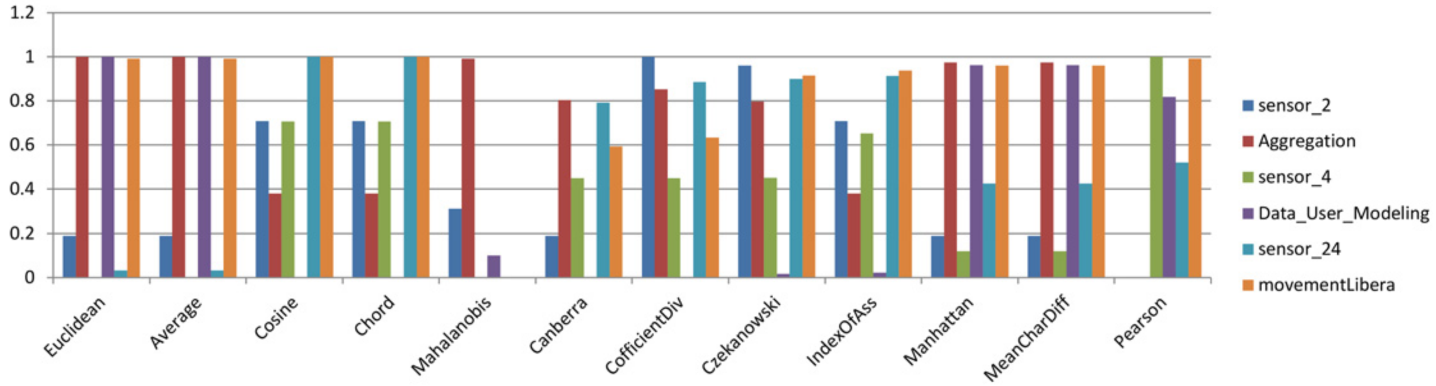
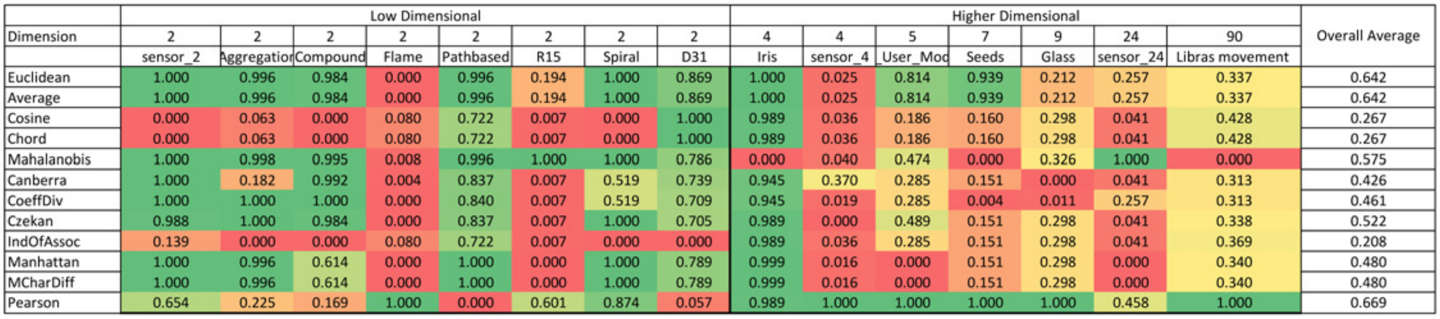


Fig 8. Bar chart of normalized Rand index values for selected datasets using the Group Average algorithm.

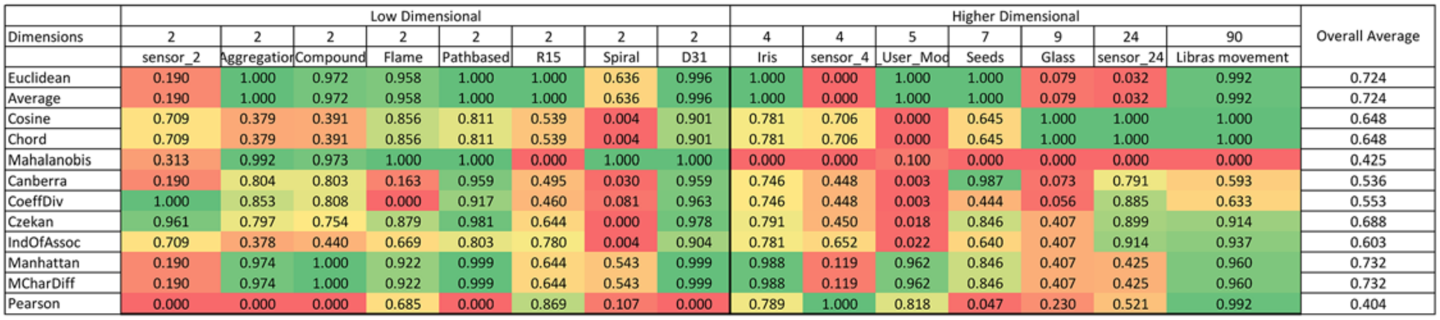
doi:10.1371/journal.pone.0144059.g008



	Low Dimensional								Higher Dimensional							Overall Average
Dimension	2	2	2	2	2	2	2	2	4	4	5	7	9	24	90	
	sensor_2	Aggregatior	Compound	Flame	Pathbased	R15	Spiral	D31	Iris	sensor_4	_User_Mod	Seeds	Glass	sensor_24	Libras movement	
Euclidean	1.000	0.996	0.984	0.000	0.996	0.194	1.000	0.869	1.000	0.025	0.814	0.939	0.212	0.257	0.337	0.642
Average	1.000	0.996	0.984	0.000	0.996	0.194	1.000	0.869	1.000	0.025	0.814	0.939	0.212	0.257	0.337	0.642
Cosine	0.000	0.063	0.000	0.080	0.722	0.007	0.000	1.000	0.989	0.036	0.186	0.160	0.298	0.041	0.428	0.267
Chord	0.000	0.063	0.000	0.080	0.722	0.007	0.000	1.000	0.989	0.036	0.186	0.160	0.298	0.041	0.428	0.267
Mahalanobis	1.000	0.998	0.995	0.008	0.996	1.000	1.000	0.786	0.000	0.040	0.474	0.000	0.326	1.000	0.000	0.575
Canberra	1.000	0.182	0.992	0.004	0.837	0.007	0.519	0.739	0.945	0.370	0.285	0.151	0.000	0.041	0.313	0.426
CoeffDiv	1.000	1.000	1.000	0.000	0.840	0.007	0.519	0.709	0.945	0.019	0.285	0.004	0.011	0.257	0.313	0.461
Czekan	0.988	1.000	0.984	0.000	0.837	0.007	1.000	0.705	0.989	0.000	0.489	0.151	0.298	0.041	0.338	0.522
IndOfAssoc	0.139	0.000	0.000	0.080	0.722	0.007	0.000	0.000	0.989	0.036	0.285	0.151	0.298	0.041	0.369	0.208
Manhattan	1.000	0.996	0.614	0.000	1.000	0.000	1.000	0.789	0.999	0.016	0.000	0.151	0.298	0.000	0.340	0.480
MCharDiff	1.000	0.996	0.614	0.000	1.000	0.000	1.000	0.789	0.999	0.016	0.000	0.151	0.298	0.000	0.340	0.480
Pearson	0.654	0.225	0.169	1.000	0.000	0.601	0.874	0.057	0.989	1.000	1.000	1.000	1.000	0.458	1.000	0.669

Fig 9. Color scale table of normalized Rand index values for the Single-link method (green is the highest and it changes color to red, which represents the lowest Rand index value).

doi:10.1371/journal.pone.0144059.g009



	Low Dimensional								Higher Dimensional							Overall Average
Dimensions	2	2	2	2	2	2	2	2	4	4	5	7	9	24	90	
	sensor_2	Aggregatior	Compound	Flame	Pathbased	R15	Spiral	D31	Iris	sensor_4	_User_Mod	Seeds	Glass	sensor_24	Libras movement	
Euclidean	0.190	1.000	0.972	0.958	1.000	1.000	0.636	0.996	1.000	0.000	1.000	1.000	0.079	0.032	0.992	0.724
Average	0.190	1.000	0.972	0.958	1.000	1.000	0.636	0.996	1.000	0.000	1.000	1.000	0.079	0.032	0.992	0.724
Cosine	0.709	0.379	0.391	0.856	0.811	0.539	0.004	0.901	0.781	0.706	0.000	0.645	1.000	1.000	1.000	0.648
Chord	0.709	0.379	0.391	0.856	0.811	0.539	0.004	0.901	0.781	0.706	0.000	0.645	1.000	1.000	1.000	0.648
Mahalanobis	0.313	0.992	0.973	1.000	1.000	0.000	1.000	1.000	0.000	0.000	0.100	0.000	0.000	0.000	0.000	0.425
Canberra	0.190	0.804	0.803	0.163	0.959	0.495	0.030	0.959	0.746	0.448	0.003	0.987	0.073	0.791	0.593	0.536
CoeffDiv	1.000	0.853	0.808	0.000	0.917	0.460	0.081	0.963	0.746	0.448	0.003	0.444	0.056	0.885	0.633	0.553
Czekan	0.961	0.797	0.754	0.879	0.981	0.644	0.000	0.978	0.791	0.450	0.018	0.846	0.407	0.899	0.914	0.688
IndOfAssoc	0.709	0.378	0.440	0.669	0.803	0.780	0.004	0.904	0.781	0.652	0.022	0.640	0.407	0.914	0.937	0.603
Manhattan	0.190	0.974	1.000	0.922	0.999	0.644	0.543	0.999	0.988	0.119	0.962	0.846	0.407	0.425	0.960	0.732
MCharDiff	0.190	0.974	1.000	0.922	0.999	0.644	0.543	0.999	0.988	0.119	0.962	0.846	0.407	0.425	0.960	0.732
Pearson	0.000	0.000	0.000	0.685	0.000	0.869	0.107	0.000	0.789	1.000	0.818	0.047	0.230	0.521	0.992	0.404

Fig 10. Color scale table of normalized Rand index values for Group Average (green is the highest and it changes color to red, which signifies the lowest Rand index value).

doi:10.1371/journal.pone.0144059.g010

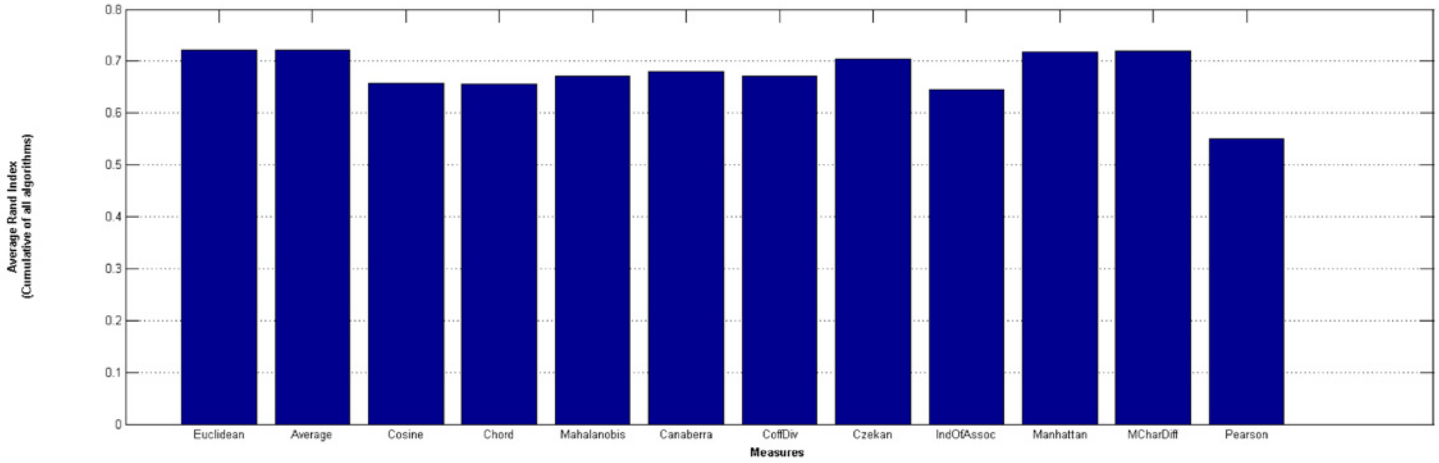


Fig 11. Overall RI Average.

doi:10.1371/journal.pone.0144059.g011

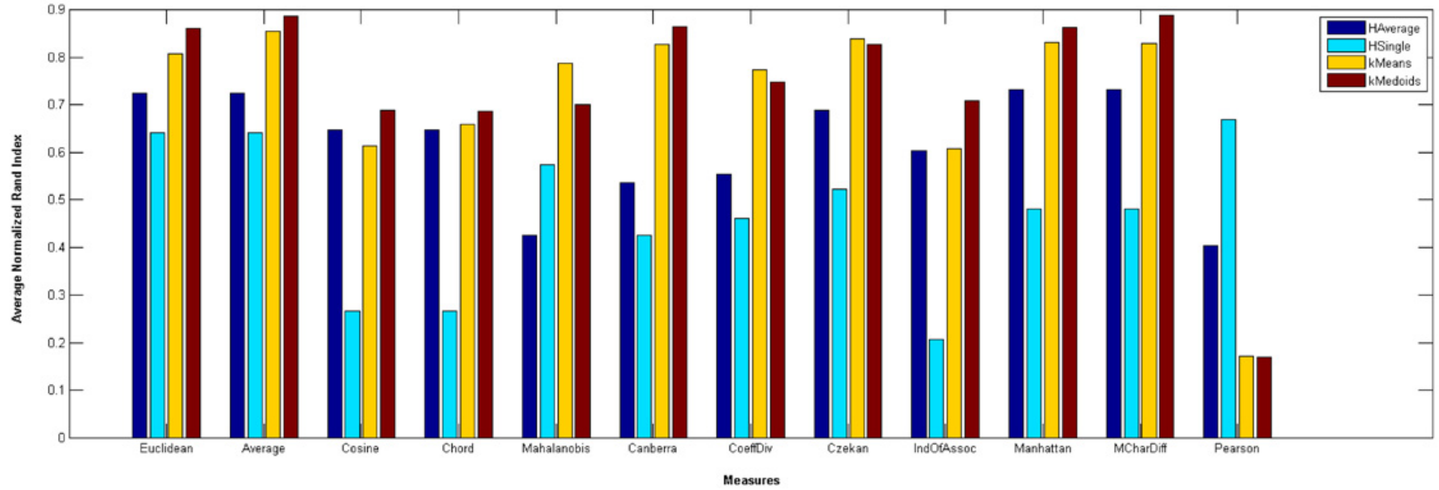


Fig 12. Average RI for four algorithms.

doi:10.1371/journal.pone.0144059.g012

Acknowledgments

Ali Seyed Shirkhorshidi would like to express his sincere gratitude to Fatemeh Zahedifar and Seyed Mohammad Reza Shirkhorshidi, who helped in revising and preparing the paper.

Author Contributions

Conceived and designed the experiments: ASS SA TYW. Performed the experiments: ASS SA TYW. Analyzed the data: ASS SA TYW. Contributed reagents/materials/analysis tools: ASS SA TYW. Wrote the paper: ASS SA TYW.

References

1. Shirkhorshidi AS, Aghabozorgi S, Wah TY, Herawan T. Big Data Clustering: A Review. Computational Science and Its Applications–ICCSA 2014. Springer; 2014. pp. 707–720. doi: 10.1007/978-3-319-09156-3_49

2. Mohebi A, Aghabozorgi S, Ying Wah T, Herawan T, Yahyapour R. Iterative big data clustering algorithms: a review. Softw Pract Exp. 2015; n/a–n/a. doi: 10.1002/spe.2341

3. Bezdek JC, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm [Internet]. Computers & Geosciences. 1984. pp. 191–203. doi: 10.1016/0098-3004(84)90020-7

4. Peters G. Some refinements of rough k-means clustering. Pattern Recognit. 2006; 39: 1481–1491. doi: 10.1016/j.patcog.2006.02.002

5. Cui W, Wang Y, Fan Y, Feng Y, Lei T. Localized FCM clustering with spatial information for medical image segmentation and bias field estimation. Int J Biomed Imaging. 2013; 2013: 930301. doi: 10.1155/2013/930301 PMID: 23997761

6. Ye J, Lazar NA, Li Y. Sparse geostatistical analysis in clustering fMRI time series. J Neurosci Methods. 2011; 199: 336–345. doi: 10.1016/j.jneumeth.2011.05.016 PMID: 21641934

7. Meyer G. Chinrungrueng F. J. Spatiotemporal clustering of fMRI time series in the spectral domain. Med Image Anal. 2004; 9: 51–68.

8. An L, Doerge RW. Dynamic Clustering of Gene Expression [Internet]. ISRN Bioinformatics. 2012. pp. 1–12. doi: 10.5402/2012/537217

9. De Souto MCP, Costa IG, de Araujo DS a, Ludermir TB, Schliep A. Clustering cancer gene expression data: a comparative study. BMC Bioinformatics. 2008; 9: 497. doi: 10.1186/1471-2105-9-497 PMID: 19038021

10. Ernst J, Nau GJ, Bar-Joseph Z. Clustering short time series gene expression data. Bioinformatics. 2005; 21: i159 –i168. doi: 10.1093/bioinformatics/bti1022 PMID: 15961453

11. Moolgavkar SH, Mcclellan RO, Dewanji A, Turim J, Georg Luebeck E, Edwards M. Time-series analyses of air pollution and mortality in the United States: A subsampling approach. Environ Health Perspect. 2013; 121: 73–78. doi: 10.1289/ehp.1104507 PMID: 23108284

12. Ignaccolo R, Ghigo S, Bande S. Functional zoning for air quality. Environ Ecol Stat. 2013; 20: 109–127. doi: 10.1007/s10651-012-0210-7

13. Carbajal-Hernández JJ, Sánchez-Fernández LP, Carrasco-Ochoa J a., Martínez-Trinidad JF. Assessment and prediction of air quality using fuzzy logic and autoregressive models. Atmos Environ. Elsevier Ltd; 2012; 60: 37–50. doi: 10.1016/j.atmosenv.2012.06.004

14. Shen W, Babushkin V, Aung Z, Woon WL. An ensemble model for day-ahead electricity demand time series forecasting. Proc fourth Int Conf Futur energy Syst—e-Energy '13. New York, New York, USA: ACM Press; 2013; 51. doi: 10.1145/2487166.2487173

15. Iglesias F, Kastner W. Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns. Energies. 2013; 6: 579–597. doi: 10.3390/en6020579

16. Wijk J Van, Selow E Van. Cluster and calendar based visualization of time series data. Proc 1999 IEEE Symp Inf Vis. IEEE Comput. Soc; 1999; 4–9. doi: 10.1109/INFVIS.1999.801851

17. Aghabozorgi S, Seyed Shirkhorshidi A, Ying Wah T. Time-series clustering–A decade review. Inf Syst. 2015; 53: 16–38. doi: 10.1016/j.is.2015.04.007

18. Bache K, Lichman M. UCI Machine Learning Repository [Internet]. 2013. Available: http://archive.ics.uci.edu/ml

19. Speech and Image Processing Unit, University of Eastern Finland [Internet]. Available: http://cs.joensuu.fi/sipu/datasets/

20. Boriah S, Chandola V, Kumar V. Similarity measures for categorical data: A comparative evaluation. In Proceedings of the eighth SIAM International Conference on Data Mining. 2008. pp. 243–254. doi: 10.1137/1.9781611972788.22

21. Lourenco F, Lobo V, Bacao F. Binary-based similarity measures for categorical data and their application in Self-Organizing Maps. 2004; 1–18.

22. Deshpande R, VanderSluis B, Myers CL. Comparison of Profile Similarity Measures for Genetic Interaction Networks. PLoS One. 2013; 8: e68664. doi: 10.1371/journal.pone.0068664 PMID: 23874711

23. Strehl A, Ghosh J, Mooney R. Impact of similarity measures on web-page clustering. Work Artif Intell Web . . .. 2000; 58–64. Available: http://www.aaai.org/Papers/Workshops/2000/WS-00-01/WS00-01-011.pdf

24. Zhang Z, Huang K, Tan T. Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes. Proceedings—International Conference on Pattern Recognition. IEEE; 2006. pp. 1135–1138. doi: 10.1109/ICPR.2006.392

25. Khalifa A Al, Haranczyk M, Holliday J. Comparison of Nonbinary Similarity Coefficients for Similarity Searching, Clustering and Compound Selection. J Chem Inf Model. 2009; 49: 1193–1201. doi: 10.1021/ci8004644 PMID: 19405526

26. Dunham MH. Data Mining Introductor and Advanced Topics. Upper Saddle River, New Jersey: Prentice Hall; 2003.
27. Gan G, Ma C, Wu J. Data Clustering theory, Algorithms, and Applications. ASASIAM Series on Statistics and Applied. Society for Industrial and Applied Mathematics; 2007.
28. Han J, Kamber M, Pei J. Data mining: concepts and techniques. Morgan Kaufmann; 2006.
29. Cha Sung-Hyuk. Comprehensive survey on distance/similarity measures between probability density functions. Int J Math Model methods Appl Sci. 2007; 1: 300–307. doi: 10.1.1.154.8446
30. Mao J, Jain AK. A self-organizing network for hyperellipsoidal clustering (HEC). IEEE Trans Neural Networks. 1996; 7: 16–29. doi: 10.1109/72.478389 PMID: 18255555
31. Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM Computing Surveys. ACM; 1999. pp. 264–323. doi: 10.1145/331499.331504
32. Wilson D, Martinez T. Improved heterogeneous distance functions. JAIR. 1997; 6: 1–34. Available: http://arxiv.org/abs/cs/9701101
33. Xu R, Wunsch D. Survey of clustering algorithms [Internet]. IEEE Transactions on Neural Networks. 2005. pp. 645–678. doi: 10.1109/TNN.2005.845141 PMID: 15940994
34. Perlibakas V. Distance measures for PCA-based face recognition. Pattern Recognit Lett. 2004; 25: 711–724. doi: 10.1016/j.patrec.2004.01.011
35. Legendre P, Legendre L. Numerical ecology. Elsevier; 2012.
36. Wang H, Wang H, Wang W, Wang W, Yang H, Yang H, et al. Clustering by pattern similarity in large data sets. 2002 ACM SIGMOD international conference on Management of Data. New York, New York, USA: ACM Press; 2002. p. 394. doi: 10.1145/564691.564737
37. Hand D, Mannila H, Smyth P. Principles of data mining(adaptive computation and machine learning). Drug safety. 2001.
38. Ji M, Xie F, Ping Y. A dynamic fuzzy cluster algorithm for time series. Abstr Appl Anal. 2013; 2013: 1–7. doi: 10.1155/2013/183410
39. János Abonyi BF. Cluster Analysis for Data Mining and System Identification. Springer; 2007.
40. Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: A survey. IEEE Trans Knowl Data Eng. 2004; 16: 1370–1386. doi: 10.1109/TKDE.2004.68
41. Santos JM, Embrechts M. On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2009. pp. 175–184. doi: 10.1007/978-3-642-04277-5_18
42. Hubert L, Arabie P. Comparing partitions. J Classif. Springer; 1985; 2: 193–218. doi: 10.1007/BF01908075
43. Fisher R. Statistical methods for research workers [Internet]. Edinburgh: Oliver and Boyd; 1925. Available: https://scholar.google.com/scholar?hl=en&q=Statistical+Methods+for+Research+Workers&btnG=&as_sdt=1%2C5&as_sdtp=#0
44. Cumming G. Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis [Internet]. 2013. Available: https://books.google.com/books?hl=en&lr=&id=1W6laNc7Xt8C&oi=fnd&pg=PR1&dq=Understanding+The+New+Statistics:+Effect+Sizes,+Confidence+Intervals,+and+Meta-Analysis&ots=PuHRVGc55O&sig=cEg6l3tSxFHlTI5dvubr1j7yMpI
45. Schlotzhauer S. Elementary statistics using JMP [Internet]. 2007. Available: https://books.google.com/books?hl=en&lr=&id=5JYM1WxGDz8C&oi=fnd&pg=PR3&dq=Elementary+Statistics+Using+JMP&ots=MZOht9zZOP&sig=IFCsAn4Nd9clwioPf3qS_QXPzKc
46. Gionis A, Mannila H, Tsaparas P. Clustering aggregation. ACM Trans Knowl Discov Data. 2005; 1: Article 4. doi: 10.1109/ICDE.2005.34
47. Zahn CT. Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters. IEEE Trans Comput. 1971; C-20: 68–86. doi: 10.1109/T-C.1971.223083
48. Veenman CJ, Reinders MJT, Backer E. A maximum variance cluster algorithm. IEEE Trans Pattern Anal Mach Intell. 2002; 24: 1273–1280. doi: 10.1109/TPAMI.2002.1033218
49. Fu L, Medico E. FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. BMC Bioinformatics. 2007; 8: 3. doi: 10.1186/1471-2105-8-3 PMID: 17204155