Effect of choice of a distance/similarity/dissimilarity metric on the performance evaluation of a clustering/classification algorithm: A Systematic Review

Coursework Group 93

Z Hussain Syed

Mohammed Waseem Uddin

Bilal Farooqi

Ananth Mohan Reddy Sudey

29 April 2022

# Introduction

The context of this work is mainly reviewing the effect of the choice of a distance/similarity/dissimilarity measure on clustering/classification algorithm performance. Clustering is a type of unsupervised learning useful method in data mining for discovering new and useful knowledge from unlabeled datasets (Vagni et al., 2021a) in which similar data or objects are collected into a single group (Nizam and Hassan, 2020a) which is applicable in various domains including medicine (Nizam and Hassan, 2020a; Vagni et al., 2021a), demographics (Abbas et al., 2020a), social network analysis (Yang et al., 2019a), etc. Clustering is based on the similarity/distance between two data objects or one object and the cluster centroid (Vagni et al., 2021a). The clustering results not only depend on the choice of the clustering algorithm but also on the type of distance/similarity/dissimilarity metric used in the process which is exemplified in (Nizam and Hassan, 2020a; Vagni et al., 2021a). Similarly, (Yan et al., 2019a) proposed a tailored image-to-subcategory distance measure for their work instead of the existing distance measures used in nearest neighbor classifiers.

In the medical domain, better clustering or categorization of patients of different biological aspects lead to better diagnosis, treatment, and prognosis (Serra et al., 2015a). (Taghva and Veni, 2010a) studied the effects of various similarity measures in the domain of document clustering. (Grua and Hoogendoorn, 2018a) applied different clustering methods with three similarity metrics: Euclidean distance, Dynamic Time Warping, and high-level features to evaluate which configuration yielded better personalization of users in the health and well-being domain. These aspects formed the motivation of this work and led to the research question “Does the choice of a dissimilarity metric effect the performance evaluation of a clustering/classification algorithm?”.

# Review Protocol

**Intervention**: different dissimilarity measures

**Population**: clustering/classification algorithms

**Comparison**: certain clustering/classification algorithms

**Outcome**: clustering/classification performance metrics

**Search** **String**: (impact OR effect) AND (comparison OR different) AND ("dissimilarity metric\*" OR "dissimilarity measure\*" OR "similarity measure\*" OR "similarity metric\*" OR "distance metric\*" OR "distance measure\*") AND (cluster OR clustering algorithm OR "clustering performance" OR "classification performance")

**Inclusion** **Criteria**:

* Include papers that talk about the impacts or effects of dissimilarity or similarity or distance measures on clustering or classification
* Include papers that address the research question in any way in a particular domain of application
* Include papers that compare different dissimilarity or similarity or distance measures
* Include papers that compare the performance of a clustering or classification algorithm if or when different dissimilarity or similarity or distance measures are used
* Included papers must be in English

**Exclusion** **Criteria**:

* Exclude opinion papers
* Exclude duplicate papers

# Results

* The number of papers identified by the search string in the initial search: 110
* The number of papers that passed the “include title” step: 62
* The number of papers that passed the “include abstract” step: 62
* The number of papers that passed the “exclude” step (the number of papers with “Yes” in the “accept contents” column): 51

# Synthesis

|  |  |
| --- | --- |
| **CODE** | **EXAMPLE QUOTATION** |
| comparison | "Fig. 1 shows the number of clusters identified for each dataset, using the four distance measures. As it emerges from the figure, changing the distance measure could significantly modify the number of identified clusters" (Vagni et al., 2021a, p. 4) |
| data-dependent | "Thus, the choice of the distance measure should not be done a-priori but evaluated according to the set of data to be analyzed and the task to be accomplished." (Vagni et al., 2021a, p. 1) |
| investigation | "This paper investigates the fundamental design of commonly used similarity metrics, and provides new insights to guide their use in practice." (Wu et al., 2015a, p. 1) |
| modified/anomalous metric | "This article proposes a novel fuzzy clustering technique utilizing weighted distance measure instead of the Euclidean distance using Grey Wolf Optimizer (GWO) as the global optimization techniques" (Achom et al., 2019a, p. 1) |
| other factors | "Finally, extensive experiments have been performed to demonstrate that by comparison with the effect caused by distance metric, the choices of fuzzy degree have a more significant effect and improvement on the performances of a FCM-based clustering algorithm" (Zhao et al., 2021a) |
| selection | "In this paper we present a new statistical approach for multidimensional data clustering based on neural networks and particularly the competitive neural network; which have shown their performance in several applications, we also compared the results of the approach according to different similarity metrics used. This comparison allows us to appropriate distance measure for clustering different kind of data, which helps the researchers to take good decision in a short time." (Ohmaid et al., 2019a, p. 7) |

|  |  |
| --- | --- |
| **THEME** | **CODES** |
| Comparing which metric works better | comparison; data-dependent; other factors |
| Modified/derived/anomalous metrics for better performance | comparison; data-dependent; modified/anomalous metric; other factors |
| Other factors influencing performance along with distance/similarity metrics | comparison; other factors |
| Selection Criteria | comparison; data-dependent; investigation; other factors; selection |

|  |  |
| --- | --- |
| **MODEL THEME** | **QUOTATIONS** |
| Distance/similarity/dissimilarity metric selection is dependent on data type and domain task | "The aim of this study is to analyze the impact of different similarity measures on the clustering results obtained for datasets containing different types of variables." (Vagni et al., 2021a, p. 2).  "Thus, the choice of the distance measure should not be done a-priori but evaluated according to the set of data to be analyzed and the task to be accomplished." (Vagni et al., 2021a, p. 1)  "After the experiments, it can be concluded that on numerical data, K-Means algorithm using Manhattan distance metric, gives higher accuracy than FCM algorithm but K-means takes higher elapsed time. Whereas using Euclidean distance metric, FCM algorithm gives higher accuracy than K-means algorithm but in this case FCM takes higher elapsed time. So, it is clear that based on different distance metric one can get different accuracy, that's why one will have to take care of distance metric also while choosing the clustering algorithm." (Nizam and Hassan, 2020a, p. 6)  "In this paper we present a new statistical approach for multidimensional data clustering based on neural networks and particularly the competitive neural network; which have shown their performance in several applications, we also compared the results of the approach according to different similarity metrics used. This comparison allows us to appropriate distance measure for clustering different kind of data, which helps the researchers to take good decision in a short time." (Ohmaid et al., 2019a, p. 7)  "Our work applies Ward’s Hierarchical Agglomerative Clustering Method to interval-valued data based on the Range Euclidean Metric as a reliable alternative to be used to uncertainty quantification from interval-valued data. The range metric use makes possible different merge points be explored in the Hierarchical Cluster Analysis methods." (S&#x00E9;rgio Galdino and Dias, 2021, p. 1) |
| Domain-based modified (or derived or anomalous) distance/similarity/dissimilarity metrics produce better clustering/classification results and other factors also influence the clustering/classification performance along with comparing (or considering) distance/similarity/dissimilarity metrics. | "When k-means clustering is used it fails to cluster data when it is non-linear, since k-means is suitable to get spherical clustering. If the data forms some shape k-means clustering algorithm is not suitable. This is proven in Figure 3." (Kavitha et al., 2016a, p. 5)  "The above fig 4 shows clustering algorithm constructs graph with local PCA bases measure, it clusters but some amount of error in clustering occurs." (Kavitha et al., 2016a, p. 5)  "Fig 5 shows the result of clustering based on the proposed algorithm. It predicts perfect clustering of non-linear data." (Kavitha et al., 2016a, p. 5)  "In the first analysis, we observed that the best configuration for the DFS method was when Pearson correlation and Euclidean distance were used. In addition, all configurations that used the Pearson Correlation obtained better results than the ones that used the Spearman Correlation and the statistical test proved to be statistically significant, for all four similarity measures." (Dantas et al., 2017a, p. 5)  "Therefore, we can state that Conf1 (Pearson - Euclidean) was the best setup for the DFS method, slightly better than Conf4." (Dantas et al., 2017a, p. 4)  "However, there are many similarity queries where Euclidean distances between raw data elements fail to capture the notion of similarity (see [ 1.51 for examples). Agrawal et a1 [2] present a more intuitive idea that two series should be considered similar if they have enough nonoverlapping time-ordered pairs of subsequences that are similar. The model allows translation and amplitude scaling. It also allows non-matching gaps in the matching subsequences." (Kalpakis et al., 2001a, p. 1)  "We propose a new distance measure using the LPC cepstral coefficients of the time-series for finding similarity between time-series models." (Kalpakis et al., 2001a, p. 8)  "We compare the clustering results obtained using LPC cepstral coefficients with those obtained using other widely used methods such as DFT, DWT, PCA, and DFT of auto-correlation of the timeseries. LPC cepstral coefficients clearly perform much better than the other methods for clustering synthetic ARIMA time-series." pg.8. "In the above example it can be seen that simple Euclidean or Manhattan distances between the model parameters will not be of use. One might think of using some other distance measures such as the maximum distance between the model parameters or the distance between the principal components of the model parameters" (Kalpakis et al., 2001a, p. 3)  " this article proposes a novel fuzzy clustering technique utilizing  weighted distance measure instead of the Euclidean distance  using Grey Wolf Optimizer (GWO) as the global optimization  techniques" pg.1. "some clustering techniques based on symmetry in [1] and point symmetry in [2], [3] have been developed.  But the problem with all these clustering techniques is that  it gets easily trapped to local optima and show considerably  slow convergence rates." (Achom et al., 2019a, p. 1)  "According to the characteristics of the historical  trajectory data, the Euclidean distance used in the  traditional K-means clustering algorithm can’t  calculate the distance between altitude profiles." pg.2. "For these defects, three  aspects as following are proposed to improve the  traditional K-means algorithm: (1) the algorithm  based on TWED is used to measure the distance  between the two profiles;" (Tang et al., 2015a, p. 3)  "In this work, we  compare and analyse the effectiveness of similarity  measures such as City Block distance, Cosine  similarity, Point symmetry distance and  Dicecoefficient to improve document clusteringwith  and without the presence of ontology" (Nadana and Shriram, 2012a, p. 1)  "We found that the Euclidean  Distance, the Cosine Similarity and the Jaccard measures have  comparable effectiveness for the partitional Arabic Documents  Clustering task for finding more coherent clusters in case we  didn’t use the stemming for the testing dataset. On the other  hand the Pearson Correlation and averaged KL Divergence are  quite similar in theirs results but there are slightly better than  the other measures in the same case" (Froud et al., 2010, p. 4)  "DDML minimizes the difference between instances with the same category and the  similarity of instances from different domains to handle the  problem of the learned distance metric suffering from high  feature distribution bias in aerial scenes."(Yan et al., 2019a, p. 16)  "Results show that when  using the Manhattan distance, smoother and shorter error  variance was not reflected in the segmentation, nor did it  provide near match with the raw image shown in Figure 1 (a).  Euclidian distance, on the other hand, displayed wider error  range, but much clearer segmentation and closer matching with  the raw image. " (Sammouda and Youssef, 2014, p. 6)  "In addition, a GFCM clustering strategy is  proposed to solve the clustering problem of using FCM under  different distance metrics and fuzzy degrees, which achieves  the aim of using objective function to control the fuzziness of  clustering results." (Zhao et al., 2021a, p. 14)  "Finally, extensive  experiments have been performed to demonstrate that by comparison with the effect caused by distance metric, the choices of  fuzzy degree have a more significant effect and improvement on  the performances of a FCM-based clustering algorithm" (Zhao et al., 2021a, p. 1)  "We also investigate the effects of assigning  different importance levels to the different similarity  measures based on the corpus characteristics." pg.1. "Our proposed ensemble similarity measure is  adaptable to suit different corpus characteristics by  assigning different weights to its contextual similarity  and its semantic relatedness components." (Ittoo and Maruster, 2009, p. 4)  "Comparing the three types of dissimilarity metrics  investigated, our results show that the base pair distance  metrics provided a higher average Sensitivity and F-measure." (Tsang and Wiese, 2009, p. 7)  "In addition, tree edit distance has shown a very strong  correlation in terms of Sensitivity. However, in our results we  did not notice an overwhelming advantage over other metrics." (Tsang and Wiese, 2009, p. 7)  "A novel ADFLICM clustering algorithm for remotely  sensed imagery classification is proposed in this paper. The  proposed algorithm is able to overcome the drawbacks of the  well-known FCM by incorporating local spatial and gray level  information. The ADFLICM is effective in removing noise  pixels and reducing the edge blurring artifact simultaneously.  This advantage is based on the definition of a new local  similarity measure, which can provide proper tradeoff between the center pixel and its neighboring pixels." (Zhang et al., 2017, pp. 11, 12)  "The variety of similarity measures can cause  difficulties in selecting a suitable measure. In this paper, we  explored set-inspired similarity measures for CFRSs, including  Fuzzy sets index, Jaccard index, Sorensen coefficient, and  Symmetric difference." (Le and Le, 2021, p. 4)  "Empirical evidence also  showed that the symmetric difference measure provides better  results than all the rest of the measures." (Le and Le, 2021, p. 5)  "Firstly;  we analyzed k-means and k-medoids algorithms’ capability  to cluster the data using different distance metrics. Secondly,  data transformation techniques including scale, range and  Yeo-Johnson are applied and the clustering accuracy of the  said clustering methods is re-calculated. It has been observed  that the results after incorporating data transformation techniques are better as compared to the mean accuracy generated  by clustering methods at alone" (Abbas et al., 2020a, p. 7)  "In addition, we want to understand how the distance metric  and the selected clustering algorithm impacts performance. We  use our three distance functions and combine these with two  commonly known clustering algorithms, namely K-Medoids  clustering [16] and Hierarchical Clustering (Agglomerative  Clustering, using the complete linkage criterion) [17]." (Grua and Hoogendoorn, 2018a, p. 4)  "In this paper, a new clustering algorithm is proposed  based on the optimum one-hop distance, which is decided by  the device electronic energy not by the topology of the  network." (Xiang et al., 2008, p. 5)  "In this paper, we propose a novel ensemble classification  approach based on variable weighting clustering. In our  approach, the different contributions of attributes are taken  into consideration by variable weighting to improve the  performance of clustering" (Ding et al., 2017, p. 6)  "We compared the impact of different similarity  measures in a framework of multi-view clustering based on hierarchical and partitional algorithms." (Serra et al., 2015a, pp. 6, 7)  "We assess the effects of data  quality issues by comparing clustering solutions from a  systematically designed set of distance measures. This task  reveals the extent to which a number of data quality issues  such as missing values, data consistency issues, and mixed  data types make it challenging to compare time series  sequences. For example, how the missing values are handled  could significantly affect or bias the “similarity” measures and therefore change the clusters derived." (Lazar et al., 2017, p. 1)  "For clustering closely similar malicious behavior of malware variants, we used a probabilistic similarity based on length-normalized log-likelihood  to measure distance between each pair of sequences." (Wang et al., 2013, p. 9)  "The presented research aims to apply clustering on  Web Documents based on HTML-DOM structure of  Webpages, where the HTML-DOM structure of each  Webpage has been represented as a string of characters,  and then applying K-means clustering on the string  representationwith four different distance measures on four  different datasets. We come to a conclusion that there is a  definite impact of structural similarity of Webpages in  clustering, which may of course vary depending upon the  webpages in the datasets used" (Sarma and Mahanta, 2019, p. 6)  "The same algorithm has been  applied with four different distance measures on four  different datasets. The clustering output in each case has  been evaluated and the results have been compared." (Sarma and Mahanta, 2019, p. 1)  "This paper proposes a marine targets spatiotemporal trajectory similarity measurement method based on multidimensional Features" (Jiang et al., 2021, p. 6)  "Our work applies Ward’s Hierarchical Agglomerative Clustering Method to interval-valued data based on the  Range Euclidean Metric as a reliable alternative to be used to  uncertainty quantification from interval-valued data. The range  metric use makes possible different merge points be explored in  the Hierarchical Cluster Analysis methods." (S&#x00E9;rgio Galdino and Dias, 2021, p. 1)  "Among the proposed distance measures, no measure consistently outperformed the others. Figure (5) shows that a  measure performing best on one corpus can be the worst  choice for another. To tackle this problem and further  improve the similarity estimation, a distance metric could  be learned by incorporating supervision [28]. " (Wrzalik and Krechel, 2019, p. 6)  "Various distance measures acting as document dissimilarity estimators have been  evaluated on five datasets. The impact of clustering parameters,  input word vectors, and inverse document frequency weighting  has been examined in our experiments. Furthermore, a comparison with document similarity estimation baselines has been  performed. We demonstrate that, on average, our approach  outperforms cosine similarity of both weighted Bag-of-Words  vectors (TF-IDF and BM25) and word embedding centroids  (Word Centroid Distance)." (Wrzalik and Krechel, 2019, p. 6)  "Considering the characteristics of  internet speech data, an unsupervised clustering method is  proposed based on the trained weighted distance that is  introduced according to the prior knowledge of human  perception. The weights trained on the validation set capture  the contribution of different acoustic features on human  perception" (Wu et al., 2014, p. 4)  "The proposed algorithm initially identifies an appropriate set of  estimation matrices by using the Dynamic Time Warping (DTW)  distance in order to measure similarities between gene expression  matrices. " (Kostadinova et al., 2012, p. 1)  "In addition, the impact of  the three methods on the quality of gene clustering is evaluated  by using k-means and k-medoids clustering algorithms and two  different cluster validation measures. " (Kostadinova et al., 2012, p. 1)  "In this paper, we compared different methods with pairwise  constraints in agglomerative hierarchical clustering. The first is  the penalty method which adds a penalty term to dissimilarity  measure if constraints are broken. The second is dissimilarity  modification method which modifies dissimilarity measure if  constraints are broken.  We have shown some numerical examples by the single linkage, an asymmetric single linkage, and the centroid method.  For the single linkage and asymmetric single linkage, the  penalty method works well, but dissimilarity modification  method does not work. This is because the single linkage uses  only the minimum dissimilarity when it calculates the dissimilarity between clusters. Moreover, the single linkage using the  penalty method often has a reversal in the dendrogram.  In contrast, both the penalty method and dissimilarity modification method work well in the centroid method. This is  because the centroid method uses dissimilarity measures of  all pairs when it calculates dissimilarity between clusters." (Takumi and Miyamoto, 2012, p. 5)  "Different similarity measures  and weighting vectors are compared for the task of journal  clustering, in an effort to estimate the ranking of academic  journals." (Su et al., 2013, p. 1)  "Three fuzzy similarity measures are used to construct  different similarity measures between journals, and five OWA  weighting vectors are compared. Experimental results on five  data sets indicate that the proposed method follows human  cognition and its ranking results are generally consistent with  the trend in the RJL that is produced by human experts" (Su et al., 2013, p. 6)  "This paper has experimentally compared three typical  regional distributions (MDM, SGM, GMM) using three  dissimilarity measures (Frobenius norm, symmetrized KL  divergence, Riemannian distance) for MSNST in texture  segmentation based on the k-means clustering method and  Graph Cuts framework. The experiments using a large  number of synthesized texture images and real natural  scene images show that the GMM distribution achieves  the superior performance in both accuracy and robustness  for discriminating complex textures with the limited user  interactions and the Riemannian measure achieves the  best discriminating power. However, the Riemannian  measure suffers from the problem of low computation  efficiency, and we would like to speed it up as our future  work, which will let us design a more discriminative and  more computationally practicable segmentation system." (Han et al., 2012, p. 5)  " We proposed social embeddedness-based  centrality indices and ADs of intra and intercluster. To facilitate knowledge sharing, we build team attribute similarity based on  product- and process-related expertise. This article was the first  research to redesign an actual PD organizational architecture by  incorporating structural and team attribute similarities from the  perspective of SNA. We also integrated the modularity index (Q)  and an improved S index to find the optimal number of clusters,  which we then use with the integrated similarity matrix as inputs  to a spectral clustering algorithm to identify clusters." (Yang et al., 2019a, p. 13)  "We introduce a number of different similarity measures which can be used to determine  the (global or partial) similarity of dimensions. The similarity of  dimensions is an important prerequisite for finding the optimal oneor two-dimensional arrangement. All variants of the dimension  arrangement problem, however, are shown to be computationally  complex problem, i.e. they are NP-complete. In our implementation  to solve the dimension arrangement problem we therefore have to  use a heuristic solution which is based on an intelligent ant system.  The experimental comparison of the sequential and similarity  arrangement clearly shows the advantage of our new approach."(Ankerst et al., 1998, p. 8)  "In order to cluster the moving targets of the video in dense result of the classifican crowded scene, according to the video regulation reticulation, important role in the it can acquire the spatio-temporal motion patterns of every importance, this paper grid by spatio-temporal gradient. Using the symmetric finish the clustering off K-L(Kullback-Leibler) divergence as a distance measure, the clustering of the spatio-temporal motion patterns could be finished. The accuracy of the clustering plays an important role in the target detection, and the K-L threshold is the key for the accuracy of the clustering. Different K-L threshold will lead to different clustering effect."(Pan et al., 2017, p. 1)  "In general, Euclidean distance and Pearson’s correlation coefficient are widely used as the distance or similarity measures  for cluster analysis of gene expression data [3]. However, as  these measures depend on the actual values of the training  data, they are very much sensitive to the noise or outlier of  the data set [3], [30]. The Euclidean distance is not effective  to reflect functional similarity such as positive and negative  correlation, interdependency as well as closeness in values.  Also, Pearson’s correlation coefficient is not robust to outliers  and it may assign a high similarity score to a pair of dissimilar  genes [3], [30]. On the other hand, information measures  such as entropy and mutual information depend only on the  probability distribution of a random variable rather than on its  actual values. Hence, the mutual information has been widely  used in gene analysis problem [31], [32], [33]. In two-way clustering framework, the mutual information is used here to  measure the similarity between two objects." (Das et al., 2008, p. 4)  "Our approach mainly consists of two steps. First, we  calculate the similarity between every pair of terms. To this  end, we propose a similarity metric based on the documents  that are tagged by the terms. Second, based on the similarity  metric, we infer a taxonomy of the terms by repeatedly  applying k-medoids clustering on the terms."(Wang et al., 2012, p. 2)  "Seemingly the similarity of fixed-length sub-series is  able to be measured with Euclidean distance or other parallel  measures. In fact, for the great differences in baseline, scale  and phase between two series, Euclidean distance is not a  sound choice. We will introduce other similarity measures  applicable to time series here. " (Xiang and Chen, 2009, p. 3)  "It has been proved that NP classifiers have an asymptotic error  rate that is at most twice the Bayes error rate, no matter the  distance used in the classification [5]. However, this advantage  decreases when the dimensionality of the input data increases  [6]. Unfortunately, as the dimensionality of the input becomes higher, the choice of the distance metric becomes more important in determining the outcome of the NP classification. A commonly used distance metric is the Euclidean distance. It is based  on the assumption that the input space is isotropic or homogeneous. This assumption is not true in many practical domains." (Fernandez and Isasi, 2008, p. 1)  "Euclidean distance and Manhattan distance are used to compare the effects of different distance metrics  on the results of clusters."(Ismail et al., 2019, p. 1)  "Not only the dice similarity measure  of our method is comparable with the state of art methods but  also is more efficient than the other complex approaches in the  perspective of simplicity, reproducibility and low computational  cost" (Imtiaz et al., 2019, p. 1)  "in the  process of classification, when the classification results of each  base classifier have similar error distribution, the final  reduction of the classification effect. Aiming at this problem, this  paper proposes a method of similarity measure of decision tree  based on confusion Matrix. "(Bingzhen et al., 2020, p. 1)  "In this study, two kinds of fuzzy metric are used in order to evaluate the effects of different metric measures. Also, inter-cluster distance approaches can be monitored to this novel procedure by changing the weights of OWA Distance." (Ulutagay and Kantarci, 2014, p. 1)  "Findings from this initial investigation indicate  that choosing a subject- vs cohort-based threshold for  estimating edge maps from connectivity matrices does not  significantly impact the map topology. In contrast, the choice of  similarity measure and non-linear relationship between  similarity and edge map sparsity may have a significant impact  on map classification and the generation of parcellation atlases" (Brooks et al., 2021, p. 1)  "K-Means cluster algorithm is employed in color space of RGB by using of two different distance metrics, and the better result is selected as the initial binarization result." (Jian et al., 2016, p. 1)  "we use  both feature-based and time-series-based distance metrics to  cluster our traces, and we quantitatively show that feature-based  clusterings segregate traces by workload just as effectively as the  more compute- and space-intensive time-series-based clusterings." (Combs et al., 2014, p. 1)  "We presented an approach of modeling the different  parameters in a cluster such as network bandwidth, network  latency, disk utilization, disk bandwidth, power utilization,  temperature, etc as a distance measure between any two  nodes in the cluster. Each of these parameters can be  obtained by interpreting logs, statistics provided by systems  like HDFS and Openflow [7]. This distance was then used  to position and relocate data blocks based on the request  pattern which in turn causes the distances measure to change  again and this process goes on to enable the distributed file  system to dynamically adapt to the changing access pattern" (Srinivasan and Varma, 2014, p. 6)  (Min et al., 2014, p. 6) |
| Selection Criteria | "The aim of this study is to analyze the impact of different similarity measures on the clustering results obtained for datasets containing different types of variables." (Vagni et al., 2021a, p. 2).  "All these findings reveal that the choice of the distance measure to be used for clustering is not a trivial task, since each metric returns very different groups of elements" (Vagni et al., 2021a, p. 5)  "Fig. 1 shows the number of clusters identified for each dataset, using the four distance measures. As it emerges from the figure, changing the distance measure could significantly modify the number of identified clusters" (Vagni et al., 2021a, p. 4)  "Thus, the choice of the distance measure should not be done a-priori but evaluated according to the set of data to be analyzed and the task to be accomplished." (Vagni et al., 2021a, p. 1)  "After the experiments, it can be concluded that on numerical data, K-Means algorithm using Manhattan distance metric, gives higher accuracy than FCM algorithm but K-means takes higher elapsed time. Whereas using Euclidean distance metric, FCM algorithm gives higher accuracy than K-means algorithm but in this case FCM takes higher elapsed time. So, it is clear that based on different distance metric one can get different accuracy, that's why one will have to take care of distance metric also while choosing the clustering algorithm." (Nizam and Hassan, 2020a, p. 6)  "In this paper we present a new statistical approach for multidimensional data clustering based on neural networks and particularly the competitive neural network; which have shown their performance in several applications, we also compared the results of the approach according to different similarity metrics used. This comparison allows us to appropriate distance measure for clustering different kind of data, which helps the researchers to take good decision in a short time." (Ohmaid et al., 2019a, p. 7)  "This paper investigates the fundamental design of commonly used similarity metrics, and provides new insights to guide their use in practice." (Wu et al., 2015a, p. 1) |

# Conclusion

From deriving and observing the model themes, the effect of the choice of a distance/similarity/dissimilarity metric on the performance of clustering/classification algorithm is dependent on the type of the data and the task being dealt with (Vagni et al., 2021a). Also, different domains of application (medicine, demographics, health, and well-being, etc.) require derived or modified or anomalous distance/similarity/dissimilarity measures to improve the clustering/classification performance of the task of that particular domain which is illustrated in the quotations from the model theme “Domain-based modified (or derived or anomalous) distance/similarity/dissimilarity metrics produce better clustering/classification results and other factors also influence the clustering/classification performance along with comparing (or considering) distance/similarity/dissimilarity metrics.”.

# References

**LIST - 1**

Abbas, S.A., Aslam, A., Rehman, A.U., Abbasi, W.A., Arif, S., Kazmi, S.Z.H., 2020. K-Means and K-Medoids: Cluster Analysis on Birth Data Collected in City Muzaffarabad, Kashmir. IEEE Access 8, 151847–151855. https://doi.org/10.1109/ACCESS.2020.3014021

Grua, E.M., Hoogendoorn, M., 2018. Exploring Clustering Techniques for Effective Reinforcement Learning based Personalization for Health and Wellbeing, in: 2018 IEEE Symposium Series on Computational Intelligence (SSCI). pp. 813–820. https://doi.org/10.1109/SSCI.2018.8628621

Nizam, T., Hassan, S.I., 2020. Exemplifying the Effects of Distance Metrics on Clustering Techniques: F-measure, Accuracy and Efficiency, in: 2020 7th International Conference on Computing for Sustainable Global Development (INDIACom). pp. 39–44. https://doi.org/10.23919/INDIACom49435.2020.9083687

Serra, A., Greco, D., Tagliaferri, R., 2015. Impact of different metrics on multi-view clustering, in: 2015 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. https://doi.org/10.1109/IJCNN.2015.7280445

Taghva, K., Veni, R., 2010. Effects of Similarity Metrics on Document Clustering, in: 2010 Seventh International Conference on Information Technology: New Generations. pp. 222–226. https://doi.org/10.1109/ITNG.2010.65

Vagni, M., Giordano, N., Balestra, G., Rosati, S., 2021. Comparison of different similarity measures in hierarchical clustering, in: 2021 IEEE International Symposium on Medical Measurements and Applications (MeMeA). pp. 1–6. https://doi.org/10.1109/MeMeA52024.2021.9478746

Yan, L., Zhu, R., Mo, N., Liu, Y., 2019. Cross-Domain Distance Metric Learning Framework With Limited Target Samples for Scene Classification of Aerial Images. IEEE Transactions on Geoscience and Remote Sensing 57, 3840–3857. https://doi.org/10.1109/TGRS.2018.2888618

Yang, Q., Yang, N., Browning, T.R., Jiang, B., Yao, T., 2019. Clustering Product Development Project Organization From the Perspective of Social Network Analysis. IEEE Transactions on Engineering Management 1–15. <https://doi.org/10.1109/TEM.2019.2939398>

**LIST - 2**

Abbas, S.A., Aslam, A., Rehman, A.U., Abbasi, W.A., Arif, S. and Kazmi, S.Z.H., 2020. K-Means and K-Medoids: Cluster Analysis on Birth Data Collected in City Muzaffarabad, Kashmir. *IEEE Access*, 8, pp.151847–151855. <https://doi.org/10.1109/ACCESS.2020.3014021>.

Achom, A., Das, R., Pakray, P. and Saha, S., 2019. Classification of Microarray Gene Expression Data using Weighted Grey Wolf Optimizer based Fuzzy Clustering. In: *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*. pp.2705–2710. <https://doi.org/10.1109/TENCON.2019.8929385>.

Ankerst, M., Berchtold, S. and Keim, D.A., 1998. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In: *Proceedings IEEE Symposium on Information Visualization (Cat. No.98TB100258)*. pp.52–60. <https://doi.org/10.1109/INFVIS.1998.729559>.

Bingzhen, Z., Xiaoming, Q., Hemeng, Y. and Zhubo, Z., 2020. A Random Forest Classification Model for Transmission Line Image Processing. In: *2020 15th International Conference on Computer Science Education (ICCSE)*. pp.613–617. <https://doi.org/10.1109/ICCSE49874.2020.9201900>.

Brooks, S.J., Parks, S.M. and Stamoulis, C., 2021. Big Data-Driven Brain Parcellation from fMRI: Impact of Cohort Heterogeneity on Functional Connectivity Maps. In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*. pp.3133–3136. <https://doi.org/10.1109/EMBC46164.2021.9630267>.

Combs, J., Nazor, J., Thysell, R., Santiago, F., Hardwick, M., Olson, L., Rivoire, S., Hsu, C.-H. and Poole, S.W., 2014. Power Signatures of High-Performance Computing Workloads. In: *2014 Energy Efficient Supercomputing Workshop*. pp.70–78. <https://doi.org/10.1109/E2SC.2014.9>.

Dantas, C.A., de O. Nunes, R., Canuto, A.M.P. and Xavier-Júnior, J.C., 2017. Investigating the Impact of Similarity Metrics in an Unsupervised-Based Feature Selection Method. In: *2017 Brazilian Conference on Intelligent Systems (BRACIS)*. pp.55–60. <https://doi.org/10.1109/BRACIS.2017.61>.

Das, C., Maji, P. and Chattopadhyay, S., 2008. A Novel Biclustering Algorithm for Discovering Value-Coherent Overlapping σ-Biclusters. In: *2008 16th International Conference on Advanced Computing and Communications*. pp.148–156. <https://doi.org/10.1109/ADCOM.2008.4760441>.

Ding, H., Zhang, P., Lu, T., Gu, H. and Gu, N., 2017. Credit scoring using ensemble classification based on variable weighting clustering. In: *2017 IEEE 21st International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. pp.509–514. <https://doi.org/10.1109/CSCWD.2017.8066746>.

Fernandez, F. and Isasi, P., 2008. Local Feature Weighting in Nearest Prototype Classification. *IEEE Transactions on Neural Networks*, 19(1), pp.40–53. <https://doi.org/10.1109/TNN.2007.902955>.

Froud, H., Benslimane, R., Lachkar, A. and Ouatik, S.A., 2010. Stemming and similarity measures for Arabic Documents Clustering. In: *2010 5th International Symposium On I/V Communications and Mobile Network*. pp.1–4. <https://doi.org/10.1109/ISVC.2010.5656417>.

Galdino, S. and Dias, J., 2021. Interval-valued Data Ward amp;#x0027;s Hierarchical Agglomerative Clustering Method: Comparison of Three Representative Merge Points. In: *2021 International Conference on Engineering and Emerging Technologies (ICEET)*. pp.1–6. <https://doi.org/10.1109/ICEET53442.2021.9659628>.

Grua, E.M. and Hoogendoorn, M., 2018. Exploring Clustering Techniques for Effective Reinforcement Learning based Personalization for Health and Wellbeing. In: *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*. pp.813–820. <https://doi.org/10.1109/SSCI.2018.8628621>.

Han, S., Zhao, Y. and Tao, W., 2012. A study of regional distributions and dissimilarity measures for multi-scale nonlinear structure tensor in texture segmentation. In: *2012 IEEE International Conference on Industrial Engineering and Engineering Management*. pp.1180–1184. <https://doi.org/10.1109/IEEM.2012.6837929>.

Imtiaz, T., Rifat, S. and Fattah, S.A., 2019. Automated Brain Tumor Segmentation from MRI Data Based on Local Region Analysis. In: *2019 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON)*. pp.107–110. <https://doi.org/10.1109/BECITHCON48839.2019.9063199>.

Ismail, Z.H., Chun, A.K.K. and Shapiai Razak, M.I., 2019. Efficient Herd – Outlier Detection in Livestock Monitoring System Based on Density – Based Spatial Clustering. *IEEE Access*, 7, pp.175062–175070. <https://doi.org/10.1109/ACCESS.2019.2952912>.

Ittoo, A. and Maruster, L., 2009. Ensemble Similarity Measures for Clustering Terms. In: *2009 WRI World Congress on Computer Science and Information Engineering*. pp.315–319. <https://doi.org/10.1109/CSIE.2009.764>.

Jian, H., Xiao-pei, L. and Qian, Z., 2016. Complex scene text binarization based on graph cut. In: *2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference*. pp.691–695. <https://doi.org/10.1109/ITNEC.2016.7560449>.

Jiang, Q., Liu, Y., Sun, S. and Tan, D., 2021. Marine Targets Spatiotemporal Trajectory Similarity Measurement Method Based on Multidimensional Features. In: *2021 IEEE International Conference on Unmanned Systems (ICUS)*. pp.852–857. <https://doi.org/10.1109/ICUS52573.2021.9641245>.

Kalpakis, K., Gada, D. and Puttagunta, V., 2001. Distance measures for effective clustering of ARIMA time-series. In: *Proceedings 2001 IEEE International Conference on Data Mining*. pp.273–280. <https://doi.org/10.1109/ICDM.2001.989529>.

Kavitha, K.R., Sandeep, S. and Praveen, P.R., 2016. Improved spectral clustering using PCA based similarity measure on different Laplacian graphs. In: *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. pp.1–6. <https://doi.org/10.1109/ICCIC.2016.7919534>.

Kostadinova, E., Boeva, V., Boneva, L. and Tsiporkova, E., 2012. An Integrative DTW-based imputation method for gene expression time series data. In: *2012 6th IEEE International Conference Intelligent Systems*. pp.258–263. <https://doi.org/10.1109/IS.2012.6335145>.

Lazar, A., Jin, L., Spurlock, C.A., Wu, K. and Sim, A., 2017. Data quality challenges with missing values and mixed types in joint sequence analysis. In: *2017 IEEE International Conference on Big Data (Big Data)*. pp.2620–2627. <https://doi.org/10.1109/BigData.2017.8258222>.

Le, Q.-H. and Le, T.-X., 2021. Exploring Set-Inspired Similarity Measures for Collaborative Filtering Recommendation. In: *2021 13th International Conference on Knowledge and Systems Engineering (KSE)*. pp.1–6. <https://doi.org/10.1109/KSE53942.2021.9648825>.

Min, L., Song, Z., Yang, X. and Zhang, J., 2014. A variable scale approach for neighbor search in point cloud data. In: *2014 International Conference on Multisensor Fusion and Information Integration for Intelligent Systems (MFI)*. pp.1–6. <https://doi.org/10.1109/MFI.2014.6997742>.

Nadana, R.T. and Shriram, R., 2012. Metrics for information retrieval: A case study. In: *International Conference on Software Engineering and Mobile Application Modelling and Development (ICSEMA 2012)*. pp.1–5. <https://doi.org/10.1049/ic.2012.0146>.

Nizam, T. and Hassan, S.I., 2020. Exemplifying the Effects of Distance Metrics on Clustering Techniques: F-measure, Accuracy and Efficiency. In: *2020 7th International Conference on Computing for Sustainable Global Development (INDIACom)*. pp.39–44. <https://doi.org/10.23919/INDIACom49435.2020.9083687>.

Ohmaid, H., Timouyas, M. and Eddarouich, S., 2019. A comparative study of similarity measurement in a new neural model of Unsupervised clustering. In: *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*. pp.1–8. <https://doi.org/10.1109/ICDS47004.2019.8942242>.

Pan, L., Jing, Z. and Ping, W.J., 2017. Kullback-Leibler threshold computing method based on clustering motion patterns. In: *4th International Conference on Smart and Sustainable City (ICSSC 2017)*. pp.1–4. <https://doi.org/10.1049/cp.2017.0108>.

Sammouda, R. and Youssef, B.B., 2014. A comparison of cluster distance metrics for the segmentation of sputum color image using unsupervised hopfield neural network classifier. In: *2014 Global Summit on Computer Information Technology (GSCIT)*. pp.1–6. <https://doi.org/10.1109/GSCIT.2014.6970130>.

Sarma, M.K. and Mahanta, A.K., 2019. Clustering of Web Documents with Structure of Webpages based on the HTML Document Object Model. In: *2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*. pp.1–6. <https://doi.org/10.1109/INCOS45849.2019.8951405>.

Serra, A., Greco, D. and Tagliaferri, R., 2015. Impact of different metrics on multi-view clustering. In: *2015 International Joint Conference on Neural Networks (IJCNN)*. pp.1–8. <https://doi.org/10.1109/IJCNN.2015.7280445>.

Srinivasan, L. and Varma, V., 2014. Cost Based Approach to Block Placement for Distributed File Systems. In: *2014 International Conference on Future Internet of Things and Cloud*. pp.132–138. <https://doi.org/10.1109/FiCloud.2014.30>.

Su, P., Shang, C. and Shen, Q., 2013. OWA aggregation of fuzzy similarity relations for journal ranking. In: *2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. pp.1–7. <https://doi.org/10.1109/FUZZ-IEEE.2013.6622376>.

Taghva, K. and Veni, R., 2010. Effects of Similarity Metrics on Document Clustering. In: *2010 Seventh International Conference on Information Technology: New Generations*. pp.222–226. <https://doi.org/10.1109/ITNG.2010.65>.

Takumi, S. and Miyamoto, S., 2012. Comparing different methods of agglomerative hierarchical clustering with pairwise constraints. In: *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems*. pp.1545–1550. <https://doi.org/10.1109/SCIS-ISIS.2012.6505021>.

Tang, X., Gu, J., Shen, Z. and Chen, P., 2015. A flight profile clustering method combining twed with K-means algorithm for 4D trajectory prediction. In: *2015 Integrated Communication, Navigation and Surveillance Conference (ICNS)*. pp.S3-1-S3-9. <https://doi.org/10.1109/ICNSURV.2015.7121260>.

Tsang, H.H. and Wiese, K.C., 2009. SARNA-ensemble-predict: The effect of different dissimilarity metrics on a novel ensemble-based RNA secondary structure prediction algorithm. In: *2009 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. pp.8–15. <https://doi.org/10.1109/CIBCB.2009.4925701>.

Ulutagay, G. and Kantarci, S., 2014. Classification with fuzzy OWA distance. In: *2014 International Conference on Fuzzy Theory and Its Applications (iFUZZY2014)*. pp.195–198. <https://doi.org/10.1109/iFUZZY.2014.7091258>.

Vagni, M., Giordano, N., Balestra, G. and Rosati, S., 2021. Comparison of different similarity measures in hierarchical clustering. In: *2021 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. pp.1–6. <https://doi.org/10.1109/MeMeA52024.2021.9478746>.

Wang, H.-T., Mao, C.-H., Wei, T.-E. and Lee, H.-M., 2013. Clustering of Similar Malware Behavior via Structural Host-Sequence Comparison. In: *2013 IEEE 37th Annual Computer Software and Applications Conference*. pp.349–358. <https://doi.org/10.1109/COMPSAC.2013.60>.

Wang, S., Lo, D. and Jiang, L., 2012. Inferring semantically related software terms and their taxonomy by leveraging collaborative tagging. In: *2012 28th IEEE International Conference on Software Maintenance (ICSM)*. pp.604–607. <https://doi.org/10.1109/ICSM.2012.6405332>.

Wrzalik, M. and Krechel, D., 2019. Balanced Word Clusters for Interpretable Document Representation. In: *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. pp.103–109. <https://doi.org/10.1109/ICDARW.2019.40089>.

Wu, X., Wu, Z., Jia, J., Meng, H., Cai, L. and Li, W., 2014. Automatic speech data clustering with human perception based weighted distance. In: *The 9th International Symposium on Chinese Spoken Language Processing*. pp.216–220. <https://doi.org/10.1109/ISCSLP.2014.6936604>.

Wu, X.-M., Li, Z. and Chang, S.-F., 2015. New insights into Laplacian similarity search. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp.1949–1957. <https://doi.org/10.1109/CVPR.2015.7298805>.

Xiang, K. and Chen, J., 2009. Fast screening algorithm for electrical alternans in ECG based on dynamical pattern recognition framework. In: *2009 7th Asian Control Conference*. pp.734–739.

Xiang, M., Shi, W., Zhang, X., Luo, Z. and Yang, X., 2008. A New Clustering Algorithm Based on the Optimum One-Hop Distance in Wireless Sensor Networks. In: *2008 International Conference on Embedded Software and Systems*. pp.35–39. <https://doi.org/10.1109/ICESS.2008.6>.

Yan, L., Zhu, R., Mo, N. and Liu, Y., 2019. Cross-Domain Distance Metric Learning Framework With Limited Target Samples for Scene Classification of Aerial Images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(6), pp.3840–3857. <https://doi.org/10.1109/TGRS.2018.2888618>.

Yang, Q., Yang, N., Browning, T.R., Jiang, B. and Yao, T., 2019. Clustering Product Development Project Organization From the Perspective of Social Network Analysis. *IEEE Transactions on Engineering Management*, pp.1–15. <https://doi.org/10.1109/TEM.2019.2939398>.

Zhang, H., Wang, Q., Shi, W. and Hao, M., 2017. A Novel Adaptive Fuzzy Local Information $C$ -Means Clustering Algorithm for Remotely Sensed Imagery Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(9), pp.5057–5068. <https://doi.org/10.1109/TGRS.2017.2702061>.

Zhao, K., Dai, Y., Jia, Z. and Ji, Y., 2021. General Fuzzy C-Means Clustering Strategy: Using Objective Function to Control Fuzziness of Clustering Results. *IEEE Transactions on Fuzzy Systems*, pp.1–1. <https://doi.org/10.1109/TFUZZ.2021.3119240>.