*Salud, Ciencia y Tecnología*. 2022;2(S2):194
doi: 10.56294/saludcyt2022194

LITERATURE REVIEW

# A Review of Data and Document Clustering pertaining to various Distance Measures

## Revisión sobre la agrupación de datos y documentos en función de varias medidas de distancia

Sumathi Subbarayan[1], Hannah Grace Gunaseelan[1]

[1]Vellore Institute of Technology Chennai, School of Advanced Sciences, Department of Mathematics. Tamil Nadu, India.

**Citar como**: Subbarayan S, Gunaseelan HG. A Review of Data and Document Clustering pertaining to various Distance Measures. Salud Cienc. Tecnol. 2022;2(S2):194. https://doi.org/10.56294/saludcyt2022194

**Submitted:** 13-11-2022 **Revised:** 08-12-2022 **Accepted:** 21-12-2022 **Published:** 31-12-2022

**Editor:** Fasi Ahamad Shaik

**ABSTRACT**

Data is being generated at an increasing rate in a variety of fields as science and technology advance. The generated data are being saved for future decision-making. Data mining is the process of extracting patterns and useful information from massive amounts of data. The distance measure, which is used to calculate how different two objects are from one another, is one such instrument. We have conducted a comprehensive survey of how the distance measures behave when employed with different algorithms. Furthermore, the effectiveness and performance of some novel similarity measures proposed by other authors are investigated.

**Keywords:** Machine Learning; Data Mining; Distance Measure; Similarity Measure and Clustering.

**Resumen**

A medida que avanzan la ciencia y la tecnología de la información, se generan datos en una gran variedad de campos a un ritmo vertiginoso. Los datos generados se archivan para futuras decisiones. La minería de datos es el proceso de tomar grandes cantidades de datos y encontrar patrones e información útil. La medida de la distancia, que se utiliza para calcular lo diferentes que son dos objetos entre sí, es uno de esos instrumentos. Hemos analizado un estudio exhaustivo de cómo actúan las medidas de distancia cuando se emplean con distintos algoritmos. Además, también se estudia la eficacia y el rendimiento de algunas de las novedosas medidas de similitud propuestas por otros autores.

**Palabras clave:** Aprendizaje Automático; Minería de Datos; Medida de Distancia; Medida de Similitud y Clustering.

## INTRODUCTION

Machine learning is the process by which a machine can learn on its own without being explicitly programmed. With a wide range of applications, such as fraud detection, product recommendations, email spam filtering, and medical diagnosis, machine learning has attracted a lot of attention in recent decades.[(1)]

Learning is the automatic discovery of previously undiscovered patterns and structures in data.[(2)] Based on the sample data, the machine learning algorithm creates a mathematical model that improves the output (prediction) of the algorithm according to previous experiences. Once the underlying patterns are found, they can be used for prediction tasks and decision-making tasks. Machine Learning algorithms are broadly classified as supervised or unsupervised learning. Techniques like regression, neural networks, classification, clustering, and decision trees were used for knowledge discovery.

Identifying similar classes of objects is known as a Clustering task.[(3)] Moreover, clustering is an unsupervised learning task. Clustering techniques can be used to determine an object space's density and sparsity. The

distribution pattern of the object and the correlations among the features of the data can also be known using clustering tasks. Agglomerative, divisive, k-means, k-medoids, and fuzzy c-means are a few examples of benchmark clustering methods. For clustering tasks, the data has to be preprocessed and relevant features are selected. The extracted features are used to measure distance/similarity values. Based on the clustering criterion, a clustering algorithm is used to cluster the data. The clusters obtained are validated using a known validation index. Then, the clusters obtained are interpreted according to domain knowledge.[4] Examples of clustering tasks are the grouping of customers based on purchase patterns, grouping of genes based on their functionality.

Finding the Distance / Similarity value between two various objects is performed by using the clustering technique.[5] The effectiveness of the distance measures is being evaluated on the spot. As a result, we want to examine the efficacy of distance measures in the current data era and learn about current research on novel measures.

A numerical value that indicates how distinct two objects are from one another is called the distance between them. The range of the distance value is 0 to 1. The greater value denotes a greater disparity between the two things.[6] We encounter three different sorts of data: mixed types, non-numeric data, and numeric data. For numerical data, distance measures such as Manhattan, Euclidean, Chebyshev, Minkowski, Mahalanobis, Canberra, and a correlation coefficient are used.[7] Non-numerical data can be converted to nominal or numerical data. Usually, raw data is not used for any of the machine learning tasks. The raw data goes through a few pre-processing steps before being analyzed.

In the following section, a literature review is made on the comparative analysis of distance measures on different datasets. This comparative study seeks to determine which distance measures can be applied to various algorithms and data types. Despite the abundance of comparative studies available, researchers used to conduct comparative analyses with various validation measures to assess the efficiency of the algorithm and the quality of clusters obtained through the algorithm with manually clustered datasets. In our study, we performed a literature review on a novel similarity / distance measure that has emerged recently.

## DEVELOPMENT

Kumar[9] studied the cluster interpretation of “chemical characteristics of soil surface data" using the k-means method. The Euclidean distance measure and the cosine distance measure are used in k-means algorithm for clustering .[10] The soil surface data was collected at the Bhanapur Micro watershed of Koppal district, Karnataka. The two and three clusters obtained by the silhouette plot were analyzed to compare the cluster quality. The results from k-means clustering using the cosine distance measure and Euclidean distance measure were compared with hierarchical clustering algorithm and dendrograms. To categorize the chemical characteristics of soil surface data, the twok-means clustering approach yields the best results. In order to visualize the cluster for different ‘k’ values the hierarchical clustering was used. The numerical result from a hierarchical clustering dendrogram using Cosine similarity was almost equal to three-cluster k-means. According to the clustering result, the cosine distance measure was a better option than the Euclidean distance measure.

A microarray is used to detect abnormalities in chromosomes. With limited research, it is not possible to point out a particular distance measure that is suitable for clustering microarray datasets. Mohammed[11] evaluated partitioning around the Medoids algorithm[12] with various distances on microarray data. The distance measures used were Euclidean, Manhattan, Minkowski, cosine, and Mahalanobis. Dunn's validity index was used to evaluate the optimal cluster solution. The microarray datasets used were taken from: Colon, Epididymal and Hematopoietic stem cells. The clustering results suggest that all the distance measures were unsuitable for clustering microarray dataset. However, an optimal cluster solution was obtained while employing Mahalanobis distance with the partitioning around medoids algorithm.

By using the common itemset mining technique, Azadani[13] proposed a unique graph-based approach for summarizing biomedical text. The proposed model works with domain-specific knowledge as a domain – independent summarizer that can recognize some text in different contexts. In this proposed method, the document is transformed into an undirected weighted graph, where the sentences serve as the vertices and the links show how similar the sentences are to one another. The similarity between the two sentences is calculated using the Jaccard similarity measure between their frequent itemset. The graphs are grouped via a “minimum spanning tree-based” clustering technique. The top-ranked sentence for each cluster was the output from the suggested summarizer model. The sentences inside each cluster were ordered based on their similarity value. The performance assessment metrics ROUGE were used to evaluate the performance of the proposed summarizer model. The dataset was obtained from an open access biomedical paper on BIOMED Central. The proposed summarizer was compared with some of the standard summarization methods such as: Lex Rank, Text Rank, Bio Chain, GraphSum, SweSum, TexLex An, Lead baseline, Auto summarize, Random baseline.[14] The proposed biomedical summarization model outperforms other summarizers. The proposed summarizer has a longer average computing time. The average memory usage of this proposed summarizer was

the lowest compared to other summarizers. To further examine the impact of user-specific settings, "10-Fold cross validation" was performed. Compared to statistical feature-based –term frequency-based, and keyword-based summarization systems– the suggested model performs effectively.

Shirkhorshidi[15] made a comparative study on "similarity and dissimilarity measure in clustering continuous data". Several clustering algorithms were employed. The distance measures were calculated using the following formulas: mean character difference, Manhattan, index of association, Czekanowski coefficient, Euclidean, average and weighted Euclidean distances. The validation measures used wewere the Rand index,[16] entropy and sum of squares of error. The dataset used were Aggregation, Compound, D31, Flame, Path based, R 15, Sensor_2, Spiral, Iris, Sensor_4, data_user_modelling, Seeds, Glass, Sensor_24 and Movement Libera. The datasets collected were of different dimensional datasets ranging from 2 to 90. ANOVA test was carried out in each algorithm, with a different distance measure, which shows the impact of similarity measure on cluster quality. To determine the efficiency of the similarity measure, the normalized Rand index values for datasets utilizing k-means, k-medoids, single link algorithm (on selected datasets) and the group average method were compared. For low-dimensional datasets, the mean character difference measure was more accurate followed by group average and Euclidean distance measure. The accuracy of the cosine measure was higher for high-dimensional datasets. Still, Pearson correlation, employed with a hierarchical approach, was advisable for high-dimensional datasets. The convergence rate of the distance measure was calculated for 100 iterations. In most datasets, Pearson followed by average distance showed the highest convergent rate. Mahalanobis distance showed good accuracy when employed with a single-link algorithm of low-dimensional data. By considering every clustering result, average distance exhibits best accuracy for all clustering algorithms and has a fast convergence rate when employed with k-means algorithm.

Kaur[17] conducted a comparison of different distance measurements for the prediction of software faults. The datasets were clustered with the k-means algorithm employed with Euclidean distance measure, Sorensen distance measure and Canberra distance measure. The datasets used were collected from NASA metrics data program namely CM1, PC1, JM1 are three projects used with requirement, code and join features. Euclidean distance measure combined with k-means clustering, used as a validation technique by the ROC Curve, generated the best results with a high chance of both detection and false alarm. In case of low budget projects, the Canberra distance measure is a good option. The Sorensen distance is suitable for fault prediction in both high and low budget projects. However, comparing other algorithms with different distance measures will aid in the improvement of software fault prediction quality.

Bouhmala[18] provided evidence of "the effectiveness of the Euclidean distance metric for the clustering problem". The k-means algorithm was used for clustering.[19] The datasets used were breast cancer and wine, retrieved from the machine learning archive webpage. Each dataset was run 50 times in order to check the consistency of the clustering results. The cluster quality was analyzed based on the purity measure. The evaluation of the cost function shows a sudden decrease as time increases. The quality of the cluster initially increased; as time passed, the quality of the cluster decreased. Overall, the Euclidean distance, which is widely used, does not represent the standard of cluster quality, making it an inappropriate metric.

The impact of similarity measures on document clustering was examined by Taghva.[20] The distance measures used were Bray-Curtis distance, Canberra distance, Euclidean distance, cosine distance, variational distance, chi-square distance, and trigonometric distance. The reuters-21578, Distribution 1,0 Test Collection[21] dataset was used. The k-means algorithm was employed with the above distance measures. The clustering results showed that chi-square works best. Canberra and Euclidean distances show average performance. The other measures, like Bray-Curtis, variational and trigonometric functions, show less coherence.

For datasets of mixed features, Prasetyo[22] made comparative analyses to obtain the best distance and dissimilarity employed with the k-prototypes clustering algorithm.[23] The datasets used were collected from "the UCI machine learning archive namely Echocardiogram, Hepatitis and Zoo datasets".[24] The distance measures Euclidean, Manhattan, Chebyshev were used for numerical data. For categorical data, the measure based on the ratio of mismatches and simple matching distances was used. The clusters obtained were evaluated by silhouette index. The clustering result demonstrates that using a combination of Euclidean distance and ratio of mismatches dissimilarity for mixed feature data in conjunction with the k-prototypes algorithm yields better clustering results.

A comparative analysis of similarity metrics for text document clustering was conducted by Huang (25). The clustering algorithm used was k-means algorithm. The similarity measures used were Euclidean, cosine, Jaccard, Pearson, and average Kullback-Leibler Divergence (KLD).[26] The validation measures used were purity and entropy. The datasets used were 20NewsGroups and WebKB from the Cluto package. The clustering results show that all measures were efficiently useful for the text document clustering task; the only exception was the Euclidean measure. Pearson correlation coefficient and the averaged KLD divergence measure are close enough to manually create categorical structure. Meanwhile, the Jaccard and Pearson coefficient measures produced more unified clusters.

The k-means clustering algorithm was experimentally studied by Gupta[27] using various distance measures like squared Euclidean, Manhattan, Minkowski, Chebyshev, Sorensen, Soergel, Kuleyuski, Canberra, Lorentzian, wave hedges, square-chi, divergence, and Clark.[28]The evaluation metrics used were accuracy, performance, and reliability. The benchmark iris dataset was used. The six different variations of the iris dataset used were original data, squared original data, standardized data, squared standardized data, logarithmic standardized data, and exponential transformation of standard data. The experimental result shows Lorentzian, squared Euclidean, Minkowski and squared chi distance / similarity measures performed well.

Furthermore, Kavitha Karun[29] made a comparative study on a few similarity measures in clustering documents. The similarity measures used were cosine similarity, Euclidean distance, correlation coefficient, and Jaccard coefficient.[26] The data sets used were available through Cluto. An incremental algorithm was used for clustering the documents. The effect of the distance measurements was analyzed based on purity, a measure to check the quality of the cluster. The clustering result shows that the Jaccard and correlation coefficients were more efficient; Euclidean distance was unsatisfactory among the four measures. The cosine similarity showed average performance.

In the cosine similarity measure,[30] "the angle between the two vectors is considered, whereas the magnitude of the vectors is not." In Euclidean distance measure, it is possible to construct many vectors having the same similarity value, as the original given vector. In order to overcome the above drawbacks, Heidarian[31] proposed a hybrid geometric approach to measure the similarity level between documents. The new similarity measure incorporating the difference between magnitudes was the triangle's area similarity (TS), the sector's area similarity (SS) and a hybrid method (TS-SS) that uses the above two methods. The datasets used were 20NewsGroup, 7 Sectors, WebKB, and Classic4. The k-means algorithm was employed for clustering the documents. For each search query, the similarity level and purity of clusters of the dataset using cosine, Euclidean, and proposed model were compared by using uniqueness of the clusters, number of booleans, minimum gapscore and Purity. The documents with higher similarity values and lower similarity values were found, based on the percentage of uniqueness.The results obtained from uniqueness shows that the proposed measure is keen to identify the similarity level. The number of boolean value counts shows the cosine similarity measure has a higher percentage of boolean rates, whereas the Euclidean distance and the proposed method show a negligible percentage of boolean rates. The purity result for measuring the quality of clusters is more efficient than the cosine similarity and Euclidean distance. The results obtained by the minimum gapscore evaluation method show no significant change, and hence it is not possible to draw any conclusions about the similarity level. Thus, the model proposed proves its perfection in clustering and in measuring similarity level on different datasets.

Hesitant Fuzzy[32] sets are helpful when dealing with hesitancy in providing references to objects in a decision-making process of clustering. For document clustering, Sahu[33] introduced a novel method using a hesitant distance/similarity measure created on fuzzy hesitant sets. The hesitant distance measures were based on the well-known Hamming distance, the Euclidean distance, the Hausdorff metric, and a variety of ordered weighted distance measures. The datasets used were 20newsgroups datasets collected from UCI KDD archive. The document collected was preprocessed, followed by indexing and feature selection. Using 50 similarity measures, the documents were clustered. The clustering result shows the percentage of documents clustered using the 50 similarity formulas. The hybrid hesitant ordered weighted Euclidean distance measure works best with 98,792 % of documents clustered.

Using document clustering and the criteria of query and content similarity, Irfan[34] proposed a method to rank web pages. With a user query,[35] the algorithm finds the keywords of the query, and the hyperlink related to the query is counted to compute tf-idf score. The documents and web pages were divided into three groups based on whether they contained all of the query terms, a few query terms (with a lower bound of half the number of keywords), or no query terms. Now the cosine similarity measure is calculated between the first cluster and the query. The rank can be found by calculating the value of the cosine similarity measure. The proposed algorithm is time saving with fewer calculations, thus reducing the time complexity as well as the complexity of the calculations.

Sahu[36] proposed "a modified cosine distance measure for document clustering using Mahout with Hadoop." The suggested model works for large datasets using Mahout with Hadoop.[37] In this modified cosine similarity measure, the distance was made to lie between 0 and 2. The updated cosine distance value was squared to reduce the distance between two locations if it is between 0 and 0,5; otherwise, it was raised. The Wikipedia article dump dataset, measuring between 142MB and 1,64GB, was the source of the clustering data. Sequence file format was used to convert the data, and Seq2Sparse uses sequence file directory data to convert it to vector format. The k-means technique was used in conjunction with modified cosine similarity. Utilizing intra-cluster and inter-cluster distance, the resulting clusters were validated. The cluster quality shows that the modified cosine distance is better than the cosine distance measure. The proposed algorithm's only disadvantage is that it takes longer to execute than k-means with cosine measure.

https://doi.org/10.56294/saludcyt2022194

"A data-dependent similarity measure technique based on attribute selection" was proposed by Deng.[38] The partitional model will incorporate the significance of attributes. The significance rate of each attribute is compared for the maximum significance rate. This maximum significance rate attribute is chosen for each partition. Initially, a random point was chosen for partition. Then the partitioning was done iteratively using the maximum significance rate attribute, until the subset was not further divisible. The smallest region covered by two instances was calculated, and the probability mass of the region was calculated by the ratio of the region covered by two instances to the total region. The average probability mass of all such partitioned regions was compared to find the similarity between two instances and was compared with the m-kNN, multi-label classification algorithm[39] and the k-NN, single-label classification algorithm.[40] The dataset used consisted of synthetic data with 20 attributes. The area under the ROC curve (AUC) was used to evaluate the performance of the algorithm. The clustering result shows that the proposed algorithm improves as the number of iterations increases. As 'k' increases, the proposed algorithm shows better performance than k-NN, m-kNN. Since the proposed algorithm works on attribute selection, it can be used for anomaly detection tasks, and it can handle high-dimensional data.

The existing measures cannot be applied to different types of data simultaneously. Collaborative filtering has also drawbacks of poor versatility and sparse data.[41] Considering the above facts, Mu[42] proposed an efficient similarity measure for collaborative filtering. To address versatility, a local similarity measure was proposed for sparse problems, the global users' similarity was estimated by computing the Hellinger Distance and Jaccard value of all ratings. The linear combination of the local and global similarity with the weight coefficient was the proposed similarity measure. The real-world datasets Jester, MovieLens, Bookcrossing and Anime were used. The recommendation experiment shows that the validation measures of precision, recall and F1-measure increase slowly, as the number of recommendations increases for the MovieLens and Anime datasets. On the Jester dataset, it shows a negative trend. In the Bookcrossing dataset, it shows a fluctuation. Overall, comparative analysis of the proposed similarity measure with other measures such as cosine, Pearson coefficient, Jaccard, CPC shows optimal performance.

Zhu[43] proposed a sqrt-cos similarity measure for information retrieval. The proposed similarity measure was based on Hellinger distance with L1-norm. The regular term frequency (tf-idf) was replaced by binary weights. The datasets used for document clustering were CSTR, Log, Reuters, WebACE and WebKB. The datasets used for query-based information retrieval were MED and NPL. The k-means algorithm was employed with Euclidean in the L2-norm and Hellinger distance in the l1 norm. The clustering accuracy and normalized mutual information show improvements for Hellinger distance. The query-based information retrieval documents[44] were evaluated using validation measures of recall and precision. The recall and precision values improve for the proposed sqrt-cos similarity measure when compared with cosine similarity.

Zhu[43] developed a unique measure of similarity based on Hellinger distances, qrt cosine similarity. Sohangir[45] examined it using the novels "Sense and Sensibility," "Pride and Prejudice," and "Wuthering Heights." Surprisingly, there is a flaw in "sqrt cosine similarity" between the two identical novels (As, similarity between two identical novels must equals one). To address this, Sohangir[45] suggested to improv sqrt cosine measure, with square root of l1 norm. The datasets used were collected from different domains CSTR, DBLP, Reuters, WebKB, 20Newsgroup. The classification methods like nearest neighbor, Naïve-Bayes and support vector machines wereThe normalized Cut algorithm, k-means, k-means clustering based on "principal component analysis",[46] and symmetric non-negative matrix factorization were employed as clustering algorithms and validated using area under the receiver operating characteristic curve. This similarity metric was used with both clustering algorithms and classification techniques. The clustering results were compared with cosine and Gaussian-based similarity measures. The clustering results suggest that the proposed similarity measure works excellently for high dimensional datasets. Further, box plot representation was given to see the outliers. The average value of the validation measures suggest that the improved sqrt cosine similarity and cosine similarity perform equally well.

Based on the expected travel time between data points within the gravitational force field, Lu[47] suggested a similarity measure. The approach was to find the similarity of two data points by adding 1 to the portion proportional to the inverse of the square of the travel time; otherwise, a constant was used in order that the singularity was avoided. Based on the similarity measure, an algorithm that represented each data point as an edge-weighted tree was proposed. Each data was allocated to a singleton cluster. The similarity between two singleton clusters was the weight of the edge connecting the two-singletons. This was a repetitive step, in which the tree was sorted in increasing order; hence, the two clusters connecting the edge were merged into one single cluster, until no edge was left out. The proposed method was compared with a potential-based hierarchical agglomerative clustering method[48] and with benchmark methods such as single linkage, complete linkage, and Ward's method. The cluster produced was validated using the Fowlkes-Mallow index. The experiment on two synthetic dataset families using FM-index shows that the cluster produced by the proposed method is of good quality. On the basis of clustering studies, the suggested model appears to be

marginally superior to single-linkage. The datasets used were yeast, wine quality (both white and red), and iris. The findings of clustering for two high-dimensional datasets, namely spambase and the USPS handwritten digit, show that the suggested method underperformed Ward's method. Thus, the proposed algorithm has limitations on high dimensional datasets. The execution time was less, proving the efficiency of the proposed algorithm.

The agglomerative clustering technique is sensitive to noise and outliers. In order to overcome these problems, Cai[49] proposed a similarity measure combining the re-construction co-efficient with pairwise distance. The similarity matrix was obtained by repetitive updating of the rows of the sparse coefficient matrix by applying a hard threshold operator. Then each column of the similarity matrix was normalized to obtain a unit l2 norm. Using the proposed similarity measure, a new agglomerative clustering algorithm was developed. Initially, singleton clusters were formed by assigning each data point to a cluster. Further, the clusters were combined based on the largest affinity between two clusters until the merging of clusters was not possible. The datasets used were wine, iris, tox-171, lung, jaffe, ORL, UMist, Palm, USPS, coil20 and coil100. The proposed clustering algorithm was compared with k-means, path integral clustering,[50] graph degree linkage[51], constrained Laplacian rank[52], and L2- graph subspace clustering.[53] The clustering accuracy (ACC) and normalized mutual information (NMI) were used to evaluate the quality of clusters. The clustering results show that the proposed algorithm outperforms the majority of the approaches. The robustness of the proposed algorithm against Gaussian noise is also vigilant.

For Authorship Identification problem, Martín-del-Campo-Rodríguez[54] proposed a weighted cosine similarity measure by redefining the cosine similarity[30] with presence-absence data value of the vector representation of the document. The dataset used for the authorship identification problem consisted of two corpora, one with long texts and the other with short texts. The effectiveness of clustering was assessed using the F-Bcubed score. Agglomerative hierarchical clustering of documents was used because each author is unique. The clusters obtained from the proposed weighted cosine similarity measure were compared with clusters obtained from the cosine similarity measure. The clustering result shows an appreciable result for long documents, whereas there is no result for short documents.

Furthermore, Grace[55] proposed a similarity measure of a bipartite graph's energy. The set of documents and the set of keywords were represented as a bipartite graph.[56] The k-means clustering algorithm was employed with the proposed distance measure. Other measures like cosine, Jaccard, Euclidean, Manhattan, Canberra and maximum distance were used in k-means algorithm for comparative analysis. The benchmark datasets like classic, WebKB and BBC were used for comparison. The Sum of Squares Within value of all benchmark distance measures was compared with the proposed measure. The clusters obtained were of better quality. The proposed distance measure is desirable for document clustering.

A similarity measure was put up by Lin[57] for the classification and clustering of text documents. According to whether a feature appeared in both, just one, or neither of the two documents, the similarity between them was determined. If the feature was present in both documents, the value of similarity decreased as the difference between the features present in both documents increased. If the feature only appeared in one document, the similarity was assigned a fixed value. A value of zero was assigned to the similarity if no feature was found in either document. With k-NN (both single and multiple), k-means clustering and hierarchical agglomerative clustering[58] were employed. The proposed similarity measure was compared with Euclidean, cosine, extended Jaccard, pairwise-adaptive and IT-Sim to check its performance. The clusters obtained were validated using accuracy and entropy values. The dataset used were WebKB, Reuters-8 and RCVI. The classification accuracy using single and multiple classification k-NN algorithms, for different values of k, for the proposed similarity measure was compared with the other similarity measures. The result shows that the efficiency of the proposed similarity measure was two times more than Euclidean. The performance of IT-Sim is superior to the suggested similarity measure. The result shows that the proposed similarity measure is significantly better than the others with respect to validation measure entropy and accuracy. For WebKB and Reuters8, the proposed similarity works faster. For the high dimensional dataset RCVI, the presence-absence feature cuts down the features which are less significant. Thus, for a high dimensional dataset the proposed similarity measure can be recommended in terms of efficiency and performance.

Finally, Eminağaoğlu[59] proposed a novel similarity measure which uses the relative difference between the two instances. The range of the proposed similarity measure lies between 0 and 1. The dataset used was TTC-3600, a benchmark Turkish dataset. The Zemberek, F5, and F7 stemmers and incremental wrapper subset selection with Naïve Bayes algorithm for feature reduction were used to form different datasets. The algorithms used were the Rocchio classifier[60] employed with cosine similarity and the proposed similarity method, k-NN (for k=1) employed with Pearson coefficient, Euclidean distance, cosine similarity and proposed summarily method. The classification results obtained show improvement based on validation measures like F-Score, precision, and recall. The proposed similarity measure can be used for document classification and takes only non-negative numerical values. As a result, the similarity measure can be improved further for

categorical attributes and negative numbers.

## CONCLUSION

This literature review helps to analyze the pros and cons of distance measurements while using different algorithms. A few observations are as follows:

1) The distance measures literally mean distance but technically it is the difference between the features,
2) The distance measure is the heart of the algorithm,
3) Selecting the appropriate similarity measure can be chosen based on the type of data and the domain of application,
4) Euclidean distance goes well with dense data and applicable for low dimensional data,
5) Jaccard and cosine similarity measure are useful for sparse data,
6) Mahalanobis distance is sensitive to outliers and, therefore, can be recommended for anomaly detection.

Our future work will be to find an improved similarity method using vector space models for text mining, text classification, summarization, and categorization.

## REFERENCES

1. Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier; 2011.

2. Tan PN, Steinbach M, Kumar V. Introduction to data mining. Pearson Education India; 2016.

3. Witten IH, Frank E, Hall MA, Pal CJ, Data Mining Working Group. Practical machine learning tools and techniques. In: Data Mining. 2005;2:4.

4. Koutroumbas K, Theodoridis S. Pattern recognition. Academic Press; 2008.

5. Murphy KP. Machine learning: a probabilistic perspective. MIT Press; 2012.

6. Jiang SY, Li X, Zheng Q. Principles and practice of data mining. Publishing House of Electronics Industry; 2013.

7. Manning C, Schutze H. Foundations of statistical natural language processing. MIT Press; 1999.

8. Santini S, Jain R. Similarity measures. IEEE Trans Pattern Anal Mach Intell. 1999;21(9):871-883.

9. Kumar DA, Kannathasan N. A study and characterization of chemical properties of soil surface data using K-means algorithm. In: 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering. IEEE; 2013. pp. 264-270.

10. Anderberg MR. Cluster analysis for applications. Probability and Mathematical Statistics: A Series of Monographs and Textbooks. Vol. 19. Academic Press; 2014.

11. Mohammed, N. N., & Abdulazeez, A. M. (2017). Evaluation of partitioning around medoids algorithm with various distances on microarray data. In 2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) (pp. 1011-1016). IEEE.

12. Arora, P., & Varshney, S. (2016). Analysis of k-means and k-medoids algorithm for big data. Procedia Computer Science, 78, 507-512.

13. Azadani, M. N., Ghadiri, N., & Davoodijam, E. (2018). Graph based biomedical text summarization: An itemset mining and sentence clustering approach. Journal of biomedical informatics, 84, 42-58.

14. Antiqueira, L., Oliveira Jr, O. N., Costa, L. d. F., & Nunes, M. d. G. V. (2009). A complex network approach to text summarization. Information Sciences, 179(5), 584-599.

15. Shirkhorshidi, A. S., Aghabozorgi, S., & Wah, T. Y. (2015). A comparison study on similarity and dissimilarity measures in clustering continuous data. PloS one, 10(12), e0144059.

16. Santos, J. M., & Embrechts, M. (2009). On the use of the adjusted rand index as a metric for evaluating

supervised classification. In International conference on artificial neural networks (pp. 175-184). Springer, Berlin, Heidelberg.

17. Kaur, D. (2014). A Comparative Study of various Distance Measures for Software fault prediction. arXiv preprint arXiv:1411.7474.

18. Bouhmala, N. (2016). How good is the euclidean distance metric for the clustering problem. In 2016 5th IIAI international congress on advanced applied informatics (IIAI-AAI) (pp. 312-315). IEEE.

19. Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis. John Wiley Sons.

20. 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). IEEE.

21. Lewis DD. Reuters 21578, Distribution 1.0 Test collection. Available at: www.daviddlewis.com/resources/testcollections/reuters21578.

22. Prasetyo H, Purwarianti A. Comparison of distance and dissimilarity measures for clustering data with mix attribute types. In: The 1st International Conference on Information Technology, Computer, and Electrical Engineering; 2014. p. 276-280.

23. Ji J, Bai T, Zhou C, Ma C, Wang Z. An improved k-prototypes clustering algorithm for mixed numeric and categorical data. Neurocomputing 2013;120:590-596.

24. Dua D, Graff C. UCI Machine Learning Repository. Available at: http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science; 2019.

25. Huang A. Similarity measures for text document clustering. In: Proceedings of the sixth new zealand computer science research student conference; 2008. p. 9-56.

26. Steinbach M, Karypis G, Kumar V. A comparison of document clustering techniques. 2000.

27. Gupta MK, Chandra P. An empirical evaluation of K-means clustering algorithm using different distance/similarity metrics. In: Proceedings of ICETIT 2019. Springer; 2020. p. 884-892.

28. Jain AK. Data clustering: 50 years beyond K-means. Pattern Recognition Letters 2010;31(8):651-666.

29. Saad SM, Kamarudin SS. Comparative analysis of similarity measures for sentence level semantic measurement of text. In: 2013 IEEE international conference on control system, computing and engineering; 2013. p. 90-94.

30. Sidorov G, Gelbukh A, Gómez-Adorno H, Pinto D. Soft similarity and soft cosine measure: Similarity of features in vector space model. Computación y Sistemas 2014;18(3):491-504.

31. Heidarian A, Dinneen MJ. A hybrid geometric approach for measuring similarity level among documents and document clustering. 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService); 2016 Mar 14-17; San Francisco, USA. p. 142-151.

32. Li D, Zeng W, Zhao Y. Note on distance measure of hesitant fuzzy sets. Information Sciences. 2015;321:103-115.

33. Sahu N, Thakur GS. Hesitant distance similarity measures for document clustering. 2011 World Congress on Information and Communication Technologies; 2011 Dec 6-8; Mumbai, India. p. 430-438.

34. Irfan S, Ghosh S. Ranking web pages using cosine similarity measure. 2019 International Conference on Computing, Power and Communication Technologies (GUCON); 2019 Dec 13-14; Gurgaon, India. p. 867-870.

35. Sedding J, Kazakov D. Wordnet-based text document clustering. Proceedings of the 3rd workshop on

RObust Methods in Analysis of Natural Language Data (ROMAND 2004); 2004. p. 104-113.

36. Sahu L, Mohan BR. An improved K-means algorithm using modified cosine distance measure for document clustering using Mahout with Hadoop. 2014 9th International Conference on Industrial and Information Systems (ICIIS); 2014 Dec 12-14; Coimbatore, India. p. 1-5.

37. White T. Hadoop: The definitive guide. O'Reilly Media, Inc; 2012.

38. Deng N, Gao Z, Niu K. A Novel Data Dependent Similarity Measure Algorithm Based on Attribute Selection. 2018 IEEE International Conference on Big Data and Smart Computing (BigComp); 2018 Jan 19-22; Hong Kong, China. p. 603-606.

39. Zhang M, Zhou Z. ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition. 2007;40(7):2038-2048.

40. Cunjia F, Wang Y, Bian H. An improved KNN text classification algorithm. Foreign Electronic Measurement Technology. 2015;12:39-43.

41. Desrosiers C, Karypis G. Solving the sparsity problem: collaborative filtering via indirect similarities. 2008.

42. Mu Y, Xiao N, Tang R, Luo L, Yin X. An efficient similarity measure for collaborative filtering. In: Procedia Computer Science. 2019;147:416-421.

43. Zhu S, Liu L, Wang Y. Information retrieval using Hellinger distance and sqrt-cos similarity. In: 2012 7th International Conference on Computer Science Education (ICCSE); 2012:925-929.

44. McCallum AK. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. 1996. Available from: http://www.cs.cmu.edu/mccallum/bow/

45. Sohangir S, Wang D. Improved sqrt-cosine similarity measurement. Journal of Big Data. 2017;4(1):1-13.

46. Vidal R, Ma Y, Sastry SS. Principal component analysis. In: Generalized principal component analysis. Springer; 2016:25-62.

47. Lu Y, Hou X, Chen X. A novel travel-time based similarity measure for hierarchical clustering. Neurocomputing. 2016;173:3-8.

48. Lu Y, Wan Y. PHA: A fast potential-based hierarchical agglomerative clustering method. Pattern Recognition. 2013;46(5):1227-1239.

49. Cai Z, Yang X, Huang T, Zhu W. A new similarity combining reconstruction coefficient with pairwise distance for agglomerative clustering. Information Sciences. 2020;508:173-182.

50. Zhang W, Zhao D, Wang X. Agglomerative clustering via maximum incremental path integral. Pattern Recognition. 2013 Nov;46(11):3056-65.

51. Zhang W, Wang X, Zhao D, Tang X. Graph degree linkage: Agglomerative clustering on a directed graph. European Conference on Computer Vision. 2012;428-441.

52. Nie F, Wang X, Jordan M, Huang H. The constrained laplacian rank algorithm for graph-based clustering. Proceedings of the AAAI Conference on Artificial Intelligence. 2016;30(1).

53. Peng X, Yu Z, Yi Z, Tang H. Constructing the L2- graph for robust subspace learning and subspace clustering. IEEE Transactions on Cybernetics. 2016 Apr;47(4):1053-66.

54. Martín-del-Campo-Rodríguez C, Sidorov G, Batyrshin I. Enhancement of performance of document clustering in the authorship identification problem with a weighted cosine similarity. Mexican International Conference on Artificial Intelligence. 2018;49-56.

55. Grace GH, Desikan K. Document clustering using a new similarity measure based on energy of a bipartite graph. Indian Journal of Science and Technology. 2010;9:40.

56. Koolen JH, Moulton V. Maximal energy bipartite graphs. Graphs and Combinatorics. 2003;19(1):131-5.

57. Lin YS, Jiang JY, Lee SJ. A similarity measure for text classification and clustering. IEEE Transactions on Knowledge and Data Engineering. 2013 Jul;26(7):1575-90.

58. Jiang JY, Cheng WH, Chiou YS, Lee SJ. A similarity measure for text processing. 2011 International Conference on Machine Learning and Cybernetics. 2011;1460-5.

59. Eminağaoğlu M, Gökşen Y. A new similarity measure for document classification and text mining. KnE Social Sciences. 2020;353-66.

60. Rocchio JJ. The smart retrieval system: Experiments in automatic document processing. Relevance Feedback in Information Retrieval. 1971;313-23.

## CONFLICTS OF INTEREST

None.

## FINANCING

None.

## AUTHORSHIP CONTRIBUTION

*Conceptualization:* Sumathi Subbarayan, Hannah Grace Gunaseelan.
*Methodology:* Sumathi Subbarayan, Hannah Grace Gunaseelan.
*Writing - Original Draft:* Sumathi Subbarayan, Hannah Grace Gunaseelan.
*Writing - Review & Editing:* Sumathi Subbarayan, Hannah Grace Gunaseelan.