Journal for Re Attach Therapy and Developmental Diversities
eISSN: 2589-7799
2023 August; 6 (1): 862-869

Performance Evaluation Of Clustering – A Hybrid Approach [K- Mean- SOM NN] In Health Care

Dhivya Devi S[1], Thilagavathy S[2*], Lakshmi G[3]

[1]Assistant Professor, Faculty of Management, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu-603203, India.
[2*]Assistant Professor, Faculty of Management, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu-603203, India.
[3]Assistant Professor, Faculty of Management, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu-603203, India.

***Corresponding Author:** S. Thilagavathy
*****Assistant Professor, Faculty of Management, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu-603203, India. Email id – thilagavathysrinivasan@gmail.com, Orcid-Id-000-0003-2633-2604

Abstract

The high multitude of medical records and an array of interventions adopted for treating a specific disease contribute to the inherent sparsity of healthcare data collection. For the effective extraction of valuable information from these vast databases, novel data analytics methodologies are required. This article demonstrates an exploratory data analysis method based on a clustering algorithm for discovering trends in healthcare data. In addition, the clustering algorithm was employed in a multi-dimensional approach to concentrate on a distinct set of data parts and recognize patient groups. One of the unsupervised learning methods, Clustering, was utilized in numerous disciplines. The present research outlines the various clustering methods, specifically current similarity measures determined by distance-based clustering. This article compares K-means, Self-organizing Mapping, and K-means- SOM Hybrid algorithm. The outcomes exhibit that the hybrid algorithm precisely clustered the data. Moreover, comparing the K means and Self-organizing Mapping, the hybrid method performs better. The hybrid model performs faster and with higher clustering accuracy. Therefore, the hybrid model is highly recommended, and it can be employed to discover the patterns in the healthcare data, gain valuable knowledge concerning the early diagnosis of diseases, and identify potentially effective treatment modalities.

Keywords: Similarity, Breast Cancer, Self-organizing Mapping, Neural Network Hybrid Algorithm.

INTRODUCTION

The healthcare sector is widely recognized as one of the world's most essential and vast industries. The healthcare sector accumulates a sheer large amount of health and medical data (Shanthi et al., 2010). It is daunting to rapidly scrutinize the received data to arrive at vital decisions regarding patient health because the data is so extensive and intricate (Haraty et al., 2015). However, these data have yet to draw any insightful conclusions due to the challenges mentioned earlier. By leveraging data mining techniques, patterns from the healthcare data can be discovered, facilitating early diagnosis and adopting appropriate treatment (Delias et al., 2015). The immense volume of healthcare data and the necessity for analytical techniques rendered data mining a fascinating field of research. Data mining enables discovering and comprehending concealed trends within a dataset, which is not feasible solely through data visualization (Ogbuabor and Ugwoke, 2018).

Data mining is a promising method for extracting highly pertinent information from data (Alsayat et al., 2016). It employs automatic data analysis methods to illustrate the association between patterns within data. During patient care, physicians and healthcare institutions acquire massive volumes of data. Nevertheless, this data is challenging to manage due to its increased dimensions, sparsity, and prospective inconsistencies. Data mining concentrates on pursuing effective and efficient algorithms for transforming immense volumes of data into meaningful knowledge. Data mining approaches are broadly differentiated into supervised and unsupervised learning (Hobbs, 2011).

The objective of supervised learning is to recognize unlabeled data based on labeled input data. In predictions, supervised learning is utilized for determining the outcome variable values. This technique can derive a model from a training data set premised on previously recognized input and output variables (Chae et al., 2011). Following that, the category label for the unidentified outcome variable can be predicted. Plenty of labeled data is required for the model's learning process in supervised learning. Unsupervised learning enables the discovery of latent patterns from unlabeled data. Unlike

supervised learning, this technique has no output data for prediction. This technique identifies patterns in a data set based on the relationships between data elements.

Clustering

The process of classifying a set of related items into distinct groups is known as clustering, segregating a set of data into subgroups; subsequently, the items in each subgroup correspond to a predetermined distance measure. Distance-based clustering is a common name for the method of grouping objects using the distance function and is a recognized procedure that has proven effective. The clusters are generated so that the distance between two data points in the same group is minimal. In contrast, the distance measure is the largest for any two data points across clusters. Based on the magnitude of clustered instances, the algorithm's performance was analyzed (Negi et al., 2021).

Similarity of data

The degree of similarity reveals the proximity of data to one another. It illustrates the similarity of the data's patterns. Clustering refers to the act of classifying analogous items on the basis of similarity (Nitwattanakul. S et al, 2013). Most applications based on distance functions like Chebychev, Mahalanobis, Spearman, Chi-Square, Euclidean distance, and others employ this similarity measure to cluster items. The clusters are determined such that the data points inside a cluster have the least distance measure, and data points athwart cluster boundaries have a high distance measure. This paper addresses several limitations associated with distance-based clustering, which our research seeks to overcome.

Types of Clustering

The following are the recognized classification of clustering algorithms.

1. Partitional Clustering
2. Density-based Clustering
3. Hierarchical clustering

Partitional Clustering

Partitional Clustering is one of the prominent categories concerning clustering. In partition clustering, the algorithm separates the data into "k" partitions based on an objective function, with each segregation representing a cluster. "The clusters are generated such that items within a cluster are "alike" to one another, while items in other clusters are "distinct." Partitional clustering techniques can be adopted when the required number of clusters is static. K-means, PAM (Partition around methods), and CLARA are well-known partitioning clustering algorithms. The K-means clustering technique is popular at the same time as quick-way clustering as it is simple to implement with limited iterations (Radha et al., 2014).

Density-Based Clustering

Whenever the density of items in the data exceeds a predetermined threshold, the algorithms of density-based clustering produce a zone resembling clusters of arbitrary shape. The DBSCAN algorithm is a prominent illustration of clustering employing a density-based technique (Ester et al, 1996).

Hierarchical Clustering

The objective of algorithms for hierarchical clustering is partitioning or integrating a particular set of data into a sequence of nested partitions. These embedded partitions may have either a bottom-up (agglomerative) or a top-down (divisive) hierarchy. In the agglomerative technique, clustering commences with one item in the data of a specific cluster and recommences by clustering the nearest pairings of clusters till all items in the dataset are grouped in conjunction into a single cluster. Furthermore, Divisive hierarchical clustering begins with all items in a particular cluster and continues dividing large clusters into small clusters till all the items in the data are segregated as clusters. The illustrations of the Hierarchical clustering technique are Cluster Using Representatives (CURE) and BIR (Balance Iterative Reducing).

METRIC

The similarity between data objects is estimated by employing Distance metrics. The primary prerequisite of metric estimation in a particular problem is recruiting a suitable distance similarity function. A metric, or a distance function, represents the distance between elements or objects within a set (Galluccio et al, 2013). In clustering algorithms, metric space serves a decisive role. This article illustrates the basic clustering algorithm of k-means employing Euclidean distance metric with an example (Singh et al, 2013).

A given distance (e.g., dissimilarity) is signified to be a metric if and only if it satisfies the following four conditions:
1- Non-negativity: d(x, y) ≥ 0, for any two distinct observations p and q.
2- Symmetry: d(x, y) = d(y, x) for all x and y.
3- Triangle Inequality: **d(**x, y) ≤ d(x, r) + d(y, q) for all x,y, r.
4- d(x, y) = 0 only if x = y.

Distance metrics, such as the k-nearest neighbor's classification algorithm, are the essential premise for classification since they evaluate the dissimilarity between items of supplied data sets.

Euclidean distance
The Euclidean distance is one of the recognized standard metrics for numeric attributes or features. It refers to the typical distance of two points in a data set. The Euclidean distance is widely applied to clustering applications, including text clustering. Euclidean distance is the renege distance metric utilized with the K-means algorithm. Euclidean distance is represented by

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

The KNN algorithm is one of the prominent classifier algorithms that benefit by employing Euclidean distance to classify data.

Manhattan Distance
This metric determines the distance between two locations within a particular city. In Manhattan distance, the distance between two locations may be expressed as the number of blocks separating them. It is the exact difference between two items' coordinates in a data.

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i| \quad (2)$$

Canberra Distance
It is a weighted form of Manhattan distance used in Clustering, like Fuzzy Clustering, classification, computer security, and ham/spam detection systems. It is more robust to outliers in contrast to the preceding metric.

$$d_{ij} = \sum_{k=1}^{n} \frac{|x_{ik} - x_{jk}|}{|x_{ik}| + |x_{jk}|} \quad (3)$$

Chebyshev Distance
The Chebyshev distance among two n-D observations or vectors is equal to the maximum absolute value of the variations allying the data samples' coordinates. In a 2-D world, the Chebyshev distance between data points can be set as the sum of absolute differences of their 2-dimensional coordinates.

$$d(\mathbf{x},\mathbf{y}) = \lim_{r\to\infty}\left(\sum_{k=1}^{n}|x_k - y_k|^r\right)^{1/r}. \quad (4)$$

Example of Clustering using a Distance Metric
K-means algorithm utilizing the Euclidean distance metric
Let X be the set of data objects, and Let V= be the set of centers.
1. Randomly choose the cluster centers.
2. Calculate the distance between each data object and cluster centers.
3. Assign data objects to the cluster center whose distance from the cluster center is the minimum of all the cluster centers.
4. Calculate the new cluster center using the equation.

5. Again, calculate the distance between each data and newly obtained cluster centers.
6. Repeat the same steps 3 to 5 times till no data object is reassigned.

Som Neural Network

Self-organizing mapping neural network (Self-organizing Map, SOM) is a clustering algorithm. It is a biologically reasonable model of the artificial neural network, which can convert the input signal of any dimension into a one-dimensional or two-dimensional discrete mapping by calculating the mapping and implementing the process in an adaptive manner (Goa, 2023).

K-Means and SOM Hybrid Algorithm

The K-Means and SOM hybrid algorithm employed in this paper is a two-stage calculation method (Zhou et al., 2010). The experimental data are clustered in the first stage with the SOM clustering algorithm. All data with similar characteristics are classified into the same category so that the sample data can be grouped into different categories, and the number of categories and the center points of each class can be obtained. In the second stage, the results of the first stage are entered into the K-Means algorithm as input values and further clustered, thus forming the final clustering results.

This research proposes a breast cancer prediction model employing the K-means algorithm. The experimental data is obtained from open source in the UCI machine learning library (Santhi et al, 2010).

i. Wisconsin breast cancer data

The data collection includes 569 instances (62.7 percent benign, 37.3 percent malignant), 32 patient features inclusive of a patient identification number, 30 tumor diagnostic information, and a tumor diagnosis outcomes record (benign and malignant).

ii. Pima Indians Diabetes

The UCI machine learning repository contains the diabetes data. The NIDDK (National Institute of Diabetes and Digestive and Kidney Diseases) owns it. The instances were chosen from a more extensive database under several restrictions. Patients are females older than 21 with Pima Indian ancestry. The class 'variable' is present in 768 instances with nine characteristics. The properties have numerical values. There are 268 instances in Class 1 and 500 in Class 0, respectively (Bruno et al, 2014).

iii.Indian Liver Patient Dataset (ILPD)

To address the research objectives of this article, the UCI Machine Learning Repository was scrutinized and downloaded a database consisting of 583 entries of the ILPD (Indian Liver Patient Dataset). The ILPD data encompasses details of 583 Indian liver patients. There are 416 records for patients with liver problems and 167 for patients with non-liver issues.

Hybrid Algorithm Experimental Process for Breast Cancer Dataset.

Step 1: Initially, the UCI machine learning library's open-source Wisconsin Breast Cancer Dataset was read. The subsequent step is to encode the categorization labels in row Y, where benign tumors are represented by 0 and malignant tumors by 1. The next step is to ensure the dataset's appropriateness post-coding before employing the dataset to train a model.

Step 2: At random, the dataset was split into independent training and test sets, with 80% of the breast cancer data samples going to X-train and Y-train and the remaining 20% going to X-test and Y-test.

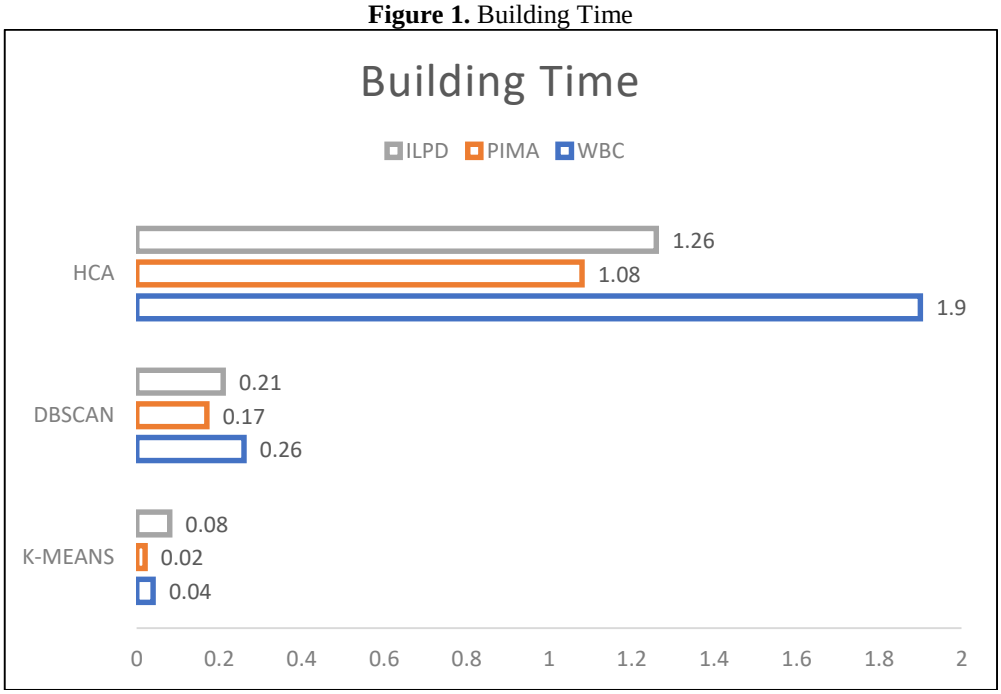
Step 3: The dataset was standardized to eliminate the negative effects brought on by varying scales amongst indicators.

Step 4: The K-Means technique was used to cluster the training set. Initially, the range from cluster number a to b was established. The elbow chart was used to determine the ideal number of clusters. The best number of clusters, as seen in Figure 1, was two, consistent with the actual outcomes. For clustering, the K Means algorithm was applied. For clustering, the hybrid K Means and SOM algorithms were also applied. The SOM neural network model's starting parameters were as follows: the input layer had 30 input nodes, and the output layer had two nodes. The beginning weight was chosen at random, the training times were 100, and theoretically, all data could be split into two categories. The K-means and SOM hybrid model's training durations were increased to 50.

Step 5: After saving the trained models, the k-means, SOM, and K-mean-SOM were tested using the test set. Finally, calculations and comparisons were made regarding each model's precision, recall, and F1 score.

Step 6: Repeat steps 1 through 5 for every other dataset.

Figure 1. Building Time



DETERMINING THE CLASSIFICATION ACCURACY

Accuracy is the degree to which the result of a measurement, calculation, or specification corresponds to the correct value or a standard, the proportion of correctly classified cases relative to the total number of cases. Table 1 displays the confusion matrix, which compares the number of accurate and inaccurate predictions made by the classification model to the actual value.

Table 1. Classification Matrix

Classification Matrix		**Actual (Target)**	
		a	**b**
Predicted Value	**a**	TP	FP
	b	FN	TN

A. True Positive Rate (TP Rate):
The proportion of accurately identified positives is measured by the true positive rate or sensitivity.

$$Sensitivit\ \ y\,(\text{TP}\ \ \text{Rate}) = \frac{TP}{TP + FN} \quad (5)$$

B. False Positive Rate (FP Rate):
The proportion of accurately identified negatives is measured by the false positive rate.

$$Specificit\ \ y\,(\text{FP}\ \ \text{Rate}) = \frac{TN}{TN + FP} \quad (6)$$

Precision and Recall Precision:
Precision and Recall Precision is a technique that examines and identifies each character with the highest level of precision.

$$\text{Precision,} \quad P = \frac{TP}{TP + FP}$$

$$\text{Recall,} \quad R = \frac{TP}{TP + FN} \qquad (8)$$

C. F- Measure:

F- Measure is a variant of accuracy not affected by negative. The following formula denotes it:

$$F = \frac{2PR}{P + R} \qquad (9)$$

D. Building Time:

Building time is the time the cluster algorithms take to make the clusters when the datasets are applied in the tool. Figure 1 illustrates the time taken by the three clustering methods, namely, K-means, DBSCAN, and HCA algorithms, to build the clusters. It is inferred from Figure 1 that the K-means algorithm had the least building time, followed by DBSCAN and HCA algorithms.

<<Figure 1>>

E. Accuracy:

Accuracy is the condition or quality of being true or correct from error. The formula for estimating accuracy is given below.

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + FN} \qquad (10)$$

Employing the above formula, the accuracy of the three clustering methods, K-means, DBSCAN, and HCA algorithms, are determined and displayed in Table 2.

Table 2. Performance Comparison of the different dataset using K-means

Dataset	Algorithm	Precision	Recall	FI-Score
	K-MEANS	0.85	0.93	0.89
	DBSCAN	0.83	0.82	0.83
WBC	HCA	0.81	0.82	0.81
	K-MEANS	0.8	0.78	0.79
	DBSCAN	0.7	0.73	0.72
PIMA	HCA	0.71	0.72	0.71
	K-MEANS	0.73	0.72	0.74
	DBSCAN	0.71	0.72	0.69
ILPA	HCA	0.69	0.68	0.66

Table 2 demonstrates that compared to the other algorithms (DBSCAN: 83% and HCA: 81%), the K-mean algorithm for WBC data achieves an accuracy rate of 85 percent, showcasing its performance efficiency. On par with the other clustering algorithms, the K-means algorithm has superior overall accuracy performance for both PIMA and ILPA datasets.

The SOM algorithm's accuracy has increased, yet it has a comparatively long run time, whereas the accuracy of the K-Means method model is high; also, it has a short run time compared to others. On the whole, the overall performance of the hybrid algorithm inclusive of K-means and SOM model is better in terms of accuracy and processing speed, making it precise for predicting breast cancer.

Table 3. Performance Comparison of Breast cancer dataset for three models

Dataset	Algorithm	Precision	Recall	FI-Score
	K-MEANS	0.85	0.93	0.89
	SOM	0.88	0.97	0.94
WBC	K Mean- SOM	0.94	0.98	0.95

CONCLUSION AND FUTURE RESEARCH DIRECTIONS

The present article initially recommends cluster analysis for PIMA and ILPA datasets and breast cancer prediction for the healthcare sector. The results of the clustering algorithms illustrated that all three methods of clustering, namely, K-means, DBSCAN, and HCA algorithms, inherited the benefits of the clustering algorithm for experimental research, but the authors recommend the k-means algorithm as it had higher accuracy and the least run time when compared to other approaches. The recommendation was made owing to the fact that the other methods took longer runtime and needed to be more accurate. Further, to overcome the shortcomings in DBSCAN and HCA algorithms, the researchers combined and incorporated the K-means algorithm with SOM and proposed a hybrid model K-means- SOM. The hybrid approach outperforms the k-means method regarding accuracy while outperforming the standard SOM method regarding accuracy and running time. Future work can validate this study by employing larger experimental datasets to evaluate the algorithms.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest: The authors declare that they have no conflict of interest.

REFERENCES

1. Aggarwal, C.C., Hinneburg, A. and Keim, D.A., 2001. On the surprising behavior of distance metrics in high dimensional space. In Database Theory—ICDT 2001: 8th International Conference London, UK, January 4–6, 2001 Proceedings 8 (pp. 420-434). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-44503-X_27
2. Alsayat, A. and El-Sayed, H., 2016, June. Efficient genetic K-Means clustering for health care knowledge discovery. In 2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA) (pp. 45-52). IEEE. https://doi.org/10.1109/SERA.2016.7516127
3. Ankerst, M., Breunig, M.M., Kriegel, H.P. and Sander, J., 1999. OPTICS: Ordering points to identify the clustering structure. ACM Sigmod record, 28(2), pp.49-60. https://doi.org/10.1145/304181.304187
4. Bruno, G., Cerquitelli, T., Chiusano, S. and Xiao, X., 2014, September. A clustering-based approach to analyse examinations for diabetic patients. In 2014 IEEE International Conference on Healthcare Informatics (pp. 45-50). IEEE. https://doi.org/10.1109/ICHI.2014.14.
5. Chae, Y.M., Kim, H.S., Tark, K.C., Park, H.J. and Ho, S.H., 2003. Analysis of healthcare quality indicator using data mining and decision support system. Expert Systems with Applications, 24(2), pp.167-172. https://doi.org/10.1016/S0957-4174(02)00139-2
6. Defays, D., 1977. An efficient algorithm for a complete link method. The computer journal, 20(4), pp.364-366. https://doi.org/10.1093/comjnl/20.4.364
7. Delias, P., Doumpos, M., Grigoroudis, E., Manolitzas, P. and Matsatsinis, N., 2015. Supporting healthcare management decisions via robust clustering of event logs. Knowledge-Based Systems, 84, pp.203-213. https://doi.org/10.1016/j.knosys.2015.04.012
8. Ester, M., Kriegel, H.P., Sander, J. and Xu, X., 1996, August. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (Vol. 96, No. 34, pp. 226-231).
9. Galluccio, L., Michel, O., Comon, P., Kliger, M. and Hero, A.O., 2013. Clustering with a new distance measure based on a dual-rooted tree. Information Sciences, 251, pp.96-113. https://doi.org/10.1016/j.ins.2013.05.040
10. Haraty, R.A., Dimishkieh, M. and Masud, M., 2015. An enhanced k-means clustering algorithm for pattern discovery in healthcare data. International Journal of Distributed Sensor Networks, 11(6), p.615740. https://doi.org/10.1155/2015/6157
11. Hobbs, G.R., 2001. Data mining and healthcare informatics. American journal of health behavior, 25(3), pp.285-289. ttps://doi.org/10.5993/AJHB.25.3.16
12. Negi, N., and Chawla, G., 2021. Clustering Algorithms in Healthcare, Intelligence Healthcare, pg 211-224.
13. Niwattanakul, S., Singthongchai, J., Naenudorn, E. and Wanapu, S., 2013, March. Using of Jaccard coefficient for keywords similarity. In Proceedings of the international multiconference of engineers and computer scientists (Vol. 1, No. 6, pp. 380-384).
14. Ogbuabor, G. and Ugwoke, F.N., 2018. Clustering algorithm for a healthcare dataset using silhouette score value. International Journal of Computer Science and Information Technology, 10(2), pp.27-37. https://doi.org/10.5121/ijcsit.2018.10203

15. Radha, R. and Rajendiran, P., 2014, February. Using K-means clustering technique to study of breast cancer. In 2014 World Congress on Computing and Communication Technologies (pp. 211-214). IEEE. https://doi.org/10.1109/WCCCT.2014.64.
16. Santhi, P. and Bhaskaran, V.M., 2010. Performance of clustering algorithms in healthcare database. International Journal for Advances in Computer Science, 2(1), pp.26-31.
17. Singh, A., Yadav, A. and Rana, A., 2013. K-means with Three different Distance Metrics. International Journal of Computer Applications, 67(10).
18. Zhou, H., Li, G.M. and Zhang, G.Y., 2010. SOM+ K-means two-phase clustering algorithm and its application. Modern Electronics Technique, 16, pp.113-115.