Journal of Soft Computing Paradigm (ISSN: 2582-2640)
www.irojournals.com/jscp/

Agglomerative Hierarchical Clustering Utilizing Principal Component Analysis with Recursive Feature Elimination for the Development of an Efficient Clustering Model

Kavitha E.[1], Anusha R.[2]

[1]Assistant Professor (SL.Grade), Department of Information Technology, University College of Engineering, Villupuram.

[2]Research Scholar, Department of ECE, Anna University, Chennai.

E-mail: [1]ekavitharesearch1@gmail.com, [2]anusharamadoss93@gmail.com

Abstract

Microarray gene expression is a technique used to monitor the expression of thousands of genes under various conditions. Clustering, an unsupervised learning technique, is employed to classify or identify similar genes by grouping sets of data objects into subclasses. This approach reveals patterns that may be obscured within extensive gene datasets and complex biological networks. Processing large-dimensional genomic datasets presents inherent complexities. To address this, the proposed method reduces the dimensionality of microarray gene datasets through a combination of feature selection and feature projection, thereby enhancing the performance of clustering algorithms. The gene datasets are processed using the Python programming language, and the output is the accuracy percentage of the validated clusters. This method has been validated using several standard datasets.

Keywords: Agglomerative Clustering, Feature Selection, Feature Projection, Model Evaluation.

DOI: https://doi.org/10.36548/jscp.2025.2.003
Received: 02.04.2025, received in revised form: 04.05.2025, accepted: 18.05.2025, published: 27.05.2025

1. Introduction

Microarray gene expression provides a view of gene activity in various biological samples. Gene expression microarrays may use genome-wide differential expression studies, disease classification, pathway analysis, gene monitoring, and more. With the selected genes, high-throughput gene expression can produce high quality data for large scale studies efficiently and economically. They can provide an effective solution for analysing the expression of known genes and transcripts. Microarrays allow monitoring of tens of thousands of genes in parallel and produce a huge amount of valuable data. The raw data is an image that must be transformed into gene expression matrices. In that matrix, the row represents genes and the columns represent values such as expression levels of a particular gene for a particular sample. In this study, the Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository is used, to distinguish malignant (cancerous) from benign (non-cancerous) samples. It contains two classes: Benign and Malignant. The main problem in most genomic datasets is their dimensionality. The problem of dimensionality reduces the performance measures of the created model.

As obtaining knowledge from any dataset with larger dimensions makes it a challenging situation to work with, several methods exist to reduce the dimensionality of the features. Here, we used Recursive Feature Elimination, which ranks the features with respect to the target value. Then, the dimensions that improve the working efficiency of the model are selected. These datasets are again subjected to the feature projection technique PCA, which further reduces the dimensions that are irrelevant to the features. Thus, we achieve a greater level of dimensionality reduction, improving the performance of clustering.

The organization of this paper is as follows. The literature on breast cancer detection and classification is presented in Section II. The proposed work is explained in detail in Section III. Experimental analysis and results are presented in Section IV. Finally, the conclusions of the proposed work are discussed in Section V.

2. Related Work

[1] Mention the noisy observations of the low Rank matrix in Sparse PCA they also seek to rework the sparsity assumptions. The above research manuscript reference is from the influential paper, which supports the principal vectors using the diagonal of the empirical

covariance. The larger entries in the diagonal are identified with high probability if $s0 \leq K1\sqrt{n}/\log p$, and they fail with high probability if $s0 \geq K2n/\log p$ for two constants $0 < K1; K2 <$ [7] proposes the identification of the most important genes and their clusters. These cluster denotethe co-expression patterns along the measured outputs using shape-based clustering models. The performed output shows the existence of similarities between the gene expression and the output variables. [4] analyse the covariance thresholding algorithm using numerical simulation technique to correctly recover the support with high probability for $s0\ K\ \sqrt{n}$ (assuming n of the same order as p), which establishes a guaranteed high-rank n much smaller than p.

[11] Used methods such as SVM, ANN, and KNN to differentiate and evaluate the cancer dataset, resulting in 66.67%. A common problem with the microarray dataset is the curse of dimensionality. When the dataset has a small number of samples, it is necessary to reduce the complexity and space. [12] focuses on the performance evaluation of grouping the dataset of E-commerce for segmentation; dimension reduction is used to transform the original data into PCA (Principal Component Analysis) features. The final output has three segments of products: that records the data in ascending order, from the least rated to the topmost, using the K-means clustering process. [13] Explains the positioning algorithms for different indoor configurations with fingerprint-based positioning algorithms; here, PCA is based on the hierarchical clustering algorithm to improve positioning accuracy through reference points (RPs) and conduct cluster-based PCA feature extraction. [14] Explains a new approach to the clustering algorithm by combining PCA and K-Means algorithms to achieve a small sample size, also reducing the interference of data dimensions. This method achieves better dimensionality reduction effects.

[15] Uses the (RFE) Recursive Feature Elimination (RFE) for feature selection, and the PCA (Principal Component Analysis) is used for feature extraction. The results confirmed that Naïve Bayesand LDA shows good performance using the combination of RFE and PCA to improve the stroke prediction.

3. Proposed Work

The proposed methodology illustrated in Figure 1 outlines a systematic process for clustering genes using the microarray method. The process commences with the loading of the dataset, specifically the Wisconsin Breast Cancer datasets from the UCI Machine Learning

Repository, to differentiate malignant (cancerous) samples from benign (non-cancerous) ones. The preprocessing phase primarily focuses on the Region of Interest (ROI) to emphasize various features, after which the system identifies and selects the area of interest to enhance the accuracy of the feature extraction method. Subsequently, feature extraction is conducted using Principal Component Analysis (PCA), followed by refinement through Recursive Feature Elimination (RFE) to cluster and select the most relevant features for this task.

Dimensionality reduction poses a significant challenge in data mining and analytics, particularly due to the large number of dimensions associated with microarray data and the processing speed required for such dimensions. To address these challenges, appropriate gene selection and the reduction of dimensions irrelevant to the class are implemented to improve the algorithm's performance. Clustering, as an unsupervised learning method, classifies distinct objects by grouping similar clusters. The agglomerative hierarchical clustering approach employs a "bottom-up" strategy, where each individual object is initially assigned its own cluster. Through the use of a similarity matrix or distance measures, pairs of clusters are merged as the process progresses up the hierarchy, ultimately forming a specified number of clusters (K). To develop an effective Agglomerative Clustering method for genomic datasets, the Recursive Feature Elimination (RFE) technique is applied to reduce features in relation to the target class, followed by the application of Principal Component Analysis (PCA). PCA serves as a feature projection technique that reduces dimensions irrelevant to the target or class values. This proposed work is depicted in Figure 1.

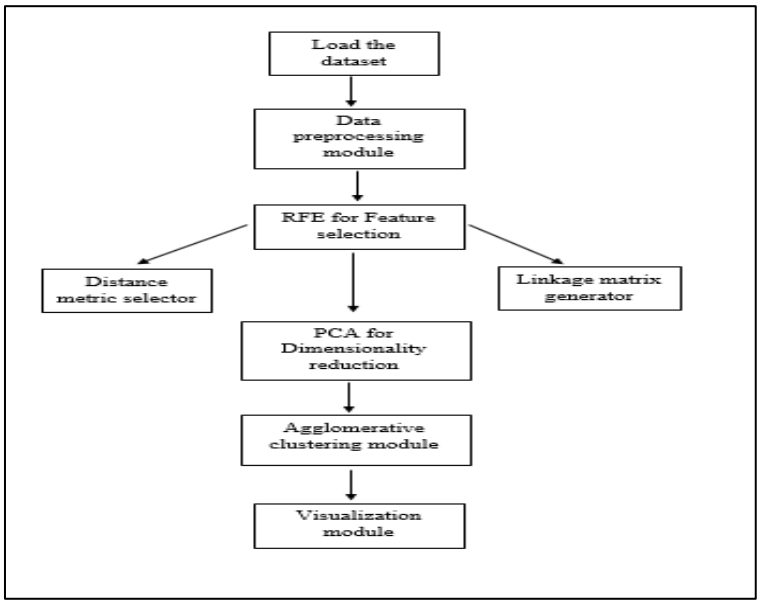

Figure 1. Flowchart of the Proposed System.

Agglomerative Clustering

Input:
i . Schemas O_1 and O_2
ii. Recalculated Similarity matrix M, between O_1 and O_2
Output: Equivalent concept pairs between O_1 and O_2
while True do
finds a pair of clusters, (a) and (b);
if(s[(a),(b)]<threshold) then
terminate the whileloop;
else
merge(a) and (b) into new cluster(a+b);
update M by deleting both the row and column corresponding to (a) and (b)
end
end
Output current clusters as equivalent concept pairs;

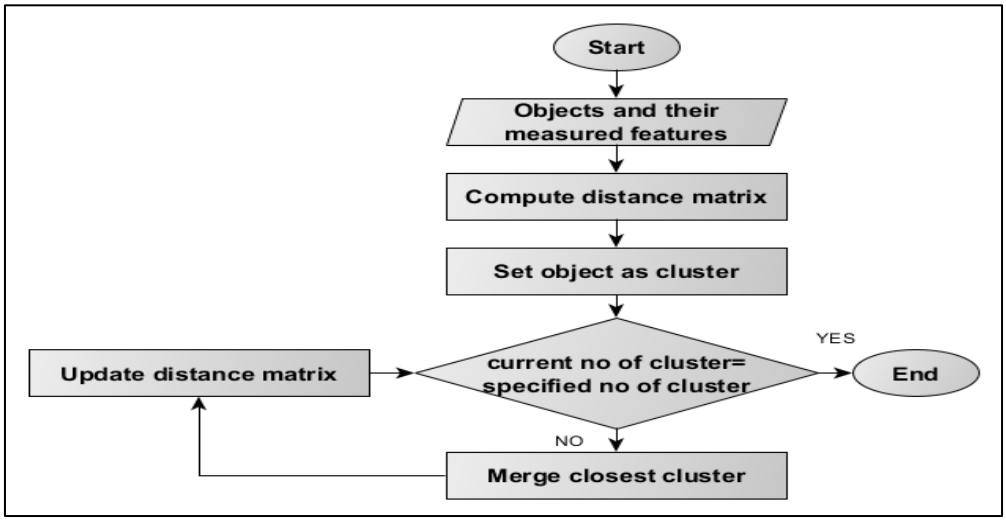

Figure 2. Working of Agglomerative Clustering Algorithm.

- **Proximity Measure**

The distance measure will determine the calculation of similarity between two elements and will influence the shape of the resulting clusters. Some of the most commonly utilized clustering methods include

Euclidean distance

$$d(x,y) = \sqrt{\sum_{i=1}^{k}(xi - yi)2}$$

- **Recursive Feature Elimination**

The goal of recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features and the importance of each feature is obtained either through a coef_ attribute or through a feature_importances_ attribute. Then, the least important features are pruned from the current set of features by using the formula.

wi = (mi(+) + mi(-)) / (si(+)+ si(-))

That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.

- **Principal Component Analysis Algorithm**

Principal Component Analysis is a method used for lowering the number of variables in a dataset. This decrease in features is also known as minimizing feature dimensionality. It involves the conversion of correlated variables into a series of uncorrelated variables to distil the genetic dataset into a unique set of attribute representations termed Eigen faces.

Step 1: Project the features in a graph.

Step 2: Calculate the average, for each gene or feature.

$$\mu = \frac{1}{M}\sum_{i=1}^{M} vi$$

Step 3: With the calculated average plot the center of the data.

Step 4: Now shift the centre of the data to the origin.

Step 5: Try to fit a line that projects through the origin, and it should fit best with the points in the graph.

Step 6: The best fit is minimizing the distance from the points to the line and maximizing the distance from the projected points to the origin.

Step 7: Thus, the best fit for the line is calculated by using the formula

$$\text{sum of squared distance} = d_1^2 + d_2^2 + d_3^2 + \cdots + d_n^2$$

Step 8: This line is called the Principal Component (PC)

Step 9: Now draw a perpendicular line to the PC's and repeat step 7.

Step 10: Calculate the variation for each PC

Variation for PC=SS (distances for PC1)/(n-1)

Step 11: By using the variation, we can get the number of dimensions we have to use for further processing.

- **Experimental Results**

In the experimental phase, we utilized samples from the Wisconsin Breast Cancer datasets in the UCI Machine Learning Repository. Certain clustering evaluation metrics are used to gain better predicted values using Accuracy Score, Mutual Information score, and Adjusted Rand Index Score. This work focuses on the target value and the predicted value, using various evaluation methods to assess the performance of clustering.

Accuracy Score

$$accuracy(\overline{y},\ y_i) = \frac{1}{\text{nsamples}} \sum_{i=0}^{\text{nsamples}} I(\overline{yi} = y_i)$$

Mutual Info Score

$$MI(U,V) = \sum_{i=1}^{|U|}\sum_{j=1}^{|V|} \frac{|U_i \cap V_j|}{N} \log \frac{N|U_i \cap V_j|}{|Ui||Vj|}$$

Adjusted Rand Index Score

ARI = (RI - Expected_RI) / (max(RI)-Expected_RI)

Model evaluation metrics using unsupervised learning are mainly for classifying distinct objects by segregating and grouping similar clusters. The focusing of the evaluation method using the aforementioned formulas of mutual information score, adjusted random score and homogeneity score is the "bottom-up" approach in the agglomerative approach. The calculation is based on each individual object that is assigned to its own cluster by analysing the similarity index or distance measure, which is then merged as one that moves up the hierarchy, forming

a new cluster. This loop continues until the K number is achieved. As illustrated in Figure 3, the comparison of the performance of hierarchical clustering algorithms with dimensionality reduction against the Breast Cancer dataset indicates that RFE-PCA agglomerative clustering outperforms both basic agglomerative clustering and PCA-agglomerative clustering on all performance measures, namely mutual information, adjusted Rand Index, and homogeneity score.

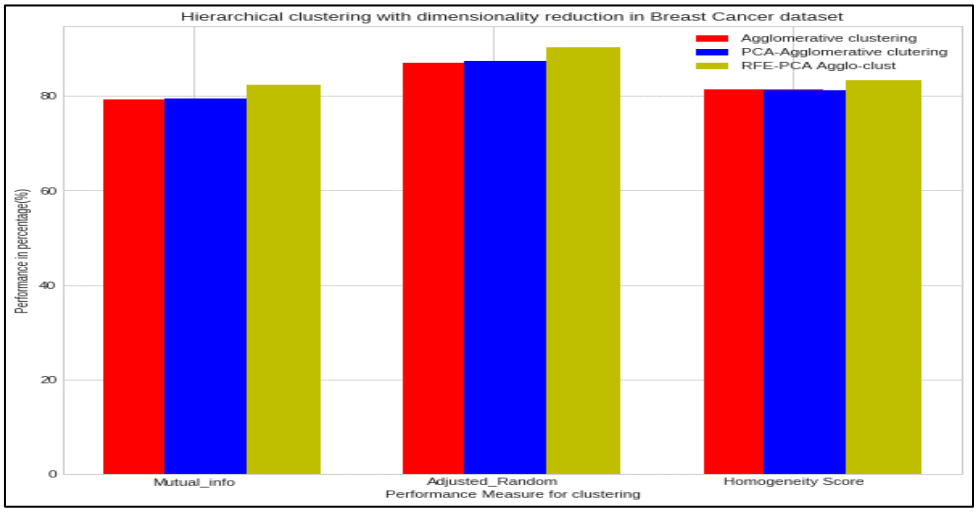

Figure 3. Performance Evaluation of Wisconsin Breast Cancer Dataset.

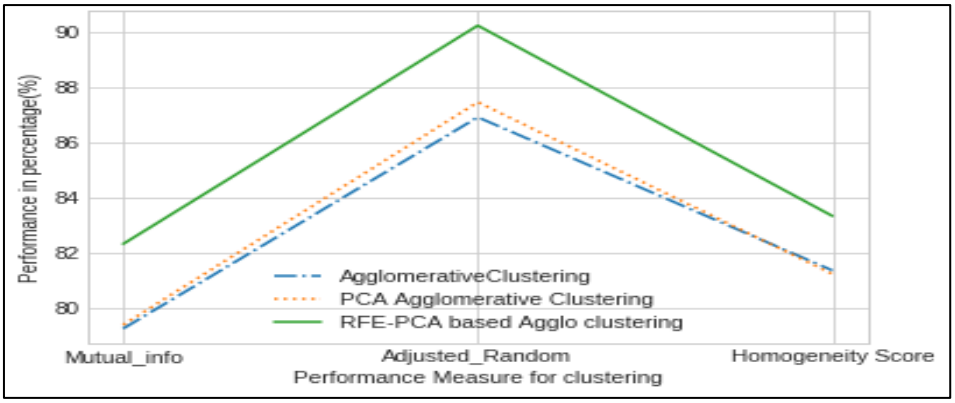

Figure 4. Evaluation of RFE-PCA Method for the Wisconsin Breast Cancer Dataset.

To highlight the trend and margin of improvement provided by dimensionality reduction approaches, especially RFE-PCA, in improving clustering accuracy, Figure 4 uses line plots. This research outcome checks the features with the targeted data and selects the relevant features. The essence of hierarchical clustering is that it is a potential tool where the dimension of the data is reduced with strategic information for a specific segmentation process that holds

substantial advantages. This manuscript not only introduce the concept of hierarchical clustering as a potent tool for preserving the original level of data with reduced dimensions.

4. Conclusion

This study proposes the integration of Recursive Feature Elimination (RFE) with Principal Component Analysis (PCA) using agglomerative clustering. The objective is to maintain the integrity of the original data within a reduced dimensional framework. This methodology employs RFE to identify pertinent features relevant to the target data, which are subsequently utilized in the PCA process to achieve significant dimensionality reduction while preserving essential feature information. Following the reduction, the processed data is analysed using the agglomerative clustering algorithm. To evaluate the performance of our approach, we implemented the existing system and assessed the effectiveness of the proposed methodology.

References

[1] Deshp, Yash, and Andrea Montanari. "Sparse PCA via covariance thresholding." Journal of Machine Learning Research 17, no. 141 (2016): 1-41.

[2] Chen, Yudong, and Jiaming Xu. "Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices." arXiv preprint arXiv:1402.1267 (2014).

[3] Liu, Wenhao, Junjun Zhai, Hongwei Ding, and Xinlong He. "The research of algorithm for protein subcellular localization prediction based on SVM-RFE." In 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), IEEE, 2017, 1-6.

[4] An, Wenjuan, Mangui Liang, and He Liu. "An improved one-class support vector machine classifier for outlier detection." Proceedings of the institution of mechanical engineers, part c: Journal of mechanical engineering science 229, no. 3 (2015): 580-588.

[5] Yongdong, Fan. "A Summary of Cross-Validation in Model Selection." Shanxi University (2013): 31-35.

[6] Zhang, S., T. Zhang, and C. Liu. "Prediction of apoptosis protein subcellular localization via heterogeneous features and hierarchical extreme learning machine." SAR and QSAR in Environmental Research 30, no. 3 (2019): 209-228.

[7] Chira, Camelia, Javier Sedano, José R. Villar, Monica Camara, and Carlos Prieto. "Shape-output gene clustering for time series microarrays." In 10th International Conference on Soft Computing Models in Industrial and Environmental Applications, pp. 241-250. Springer International Publishing, 2015.

[8] Peng, Peter, Omer Addam, Mohamad Elzohbi, Sibel T. Özyer, Ahmad Elhajj, Shang Gao, Yimin Liu et al. "Reporting and analyzing alternative clustering solutions by employing multi-objective genetic algorithm and conducting experiments on cancer data." Knowledge-Based Systems 56 (2014): 108-122.

[9] Alkhateeb, Abed, Iman Rezaeian, Siva Singireddy, and Luis Rueda. "Obtaining biomarkers in cancer progression from outliers of time-series clusters." In 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2015, 889-896.

[10] Kuo, Ren-Jieh, Y. D. Huang, Chih-Chieh Lin, Yung-Hung Wu, and Ferani E. Zulvia. "Automatic kernel clustering with bee colony optimization algorithm." Information Sciences 283 (2014): 107-122.

[11] Aldryan, D. P., and Aditsania Annisa. "Cancer detection based on microarray data classification with ant colony optimization and modified backpropagation conjugate gradient Polak-Ribiére." In 2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA), IEEE, 2018, 13-16.

[12] Valdiviezo-Diaz, Priscila. "Partitional clustering based on PCA method for segmentation of products." In 2021 16th Iberian Conference on Information Systems and Technologies (CISTI), IEEE, 2021, 1-4.

[13] Li, Ang, Jingqi Fu, Huaming Shen, and Sizhou Sun. "A cluster-principal-component-analysis-based indoor positioning algorithm." IEEE Internet of Things Journal 8, no. 1 (2020): 187-196.

[14] Huang, Jiale, Jingtong Dai, and Yanjin Li. "Research on PCA-Kmeans++ clustering algorithm considering Spatiotemporal dimension." In 2023 2nd International Conference on 3D Immersion, Interaction and Multi-sensory Experiences (ICDIIME), IEEE, 2023, 195-201.

[15] Hermiati, Arya Syifa, Rudy Herteno, Fatma Indriani, Triando Hamonangan Saragih, and Triwiyanto Triwiyanto. "A Comparative Study: Application of Principal Component Analysis and Recursive Feature Elimination in Machine Learning for Stroke Prediction." Journal of Electronics, Electromedical Engineering, and Medical Informatics 6, no. 3 (2024): 231-242.