Gene Expression Analysis using Clustering

Mentor- Dr. Ramalingaswamy Cheruku
Assistant Professor
Department of Computer Science And Engineering
National Institute of Technology, Warangal, Telangana, INDIA

Goutham Kundena
202011034

***Abstract*—To grasp complex biological processes and diseases fully. It is crucial to analyze gene expression data effectively. The rapid development of high throughput technologies like next generation sequencing has resulted in vast amounts of such data being available today. However. Extracting meaningful insights from these datasets remains a substantial challenge that warrants careful consideration of preprocessing. Feature selection, validation, and interpretation to ensure reliable and biologically valid results. Cluster analysis can greatly enhance our understanding of gene regulation by providing valuable insights. In this studys proposed method we employ an essential clustering algorithm called K means together with Particle Swarm Optimization (PSO) clustering . For analyzing gene expression data comprehensively. Our study demonstrates how PSO can efficiently determine cluster centroids based on user specified criteria while also utilizing K means clustering for seeding the initial swarm. Through evaluation using multiple gene expression datasets we compare the algorithms performance to that of K means clustering.**

***Index Terms*—K-means, PSO, gene-expression, centroids**

I. Introduction

Clustering refers to the task of grouping similar data points together based on their features or characteristics. It is an unsupervised learning technique, meaning it does not require labeled data for training. Clustering algorithms have been applied to a wide range of problems, including exploratory data analysis, data mining , image segmentation and mathematical programming. Clustering techniques have been used successfully to address the scalability problem of machine learning and data mining algorithms

The data samples are unlabelled in the real application and have a different distribution. It is challenging to categorize them into meaningful clusters. numerous clustering algorithms have emerged, acknowledging and grappling with these challenges. Hierarchical clustering (HC), with its ability to arrange data into a hierarchical structure based on the proximity matrix, yields results in the form of a binary tree or dendrogram . HC has found remarkable success in the domains of neuroimaging and bioinformatics .Partition-based clustering is well known for its simplicity and efficiency in large-scale clustering datasets, in which the dataset is divided into several subsets. The major problem with these algorithms remains the same, i.e.,all features are treated equally important during the clustering process and easily affected by the outliers, and difficult to find a meaningful cluster in the dataset.The major problem with these algorithms remains the same, i.e., all features are treated equally important during the clustering process and easily affected by the outliers, and difficult to find a meaningful cluster in the dataset.Later many weighted clustering algorithms were developed using variable weights which are multiplied with the dissimilarity measure.

The rest of the paper is organized as follows: Section 3 presents an overview of PSO and the PSO clustering technique is discussed.Preprocessing is summarized in section 4 and Experimental results are summarized in section 5.

II. Problem Statement

The objective of this study is to analyse gene expression data using various clustering methods. This research work would help the doctors to identify the stages of cancer and also enhances the medical care. This work is very convenient to avoid unnecessary biopsy.Ultimate goal of this research work is to explore various types of clustering algorithms which are suitable for analysis of cancer data

III. Proposed Methodology

A. K-means clustering

The k-Means algorithm is one of the simplest unsupervised learning algorithms. The procedure follows a simple method to classify a given data set through a certain number of clusters (assume k clusters) static a priori. The k-Means algorithm can be run multiple times to decrease the complexity of grouping data. The k-Means is a simple algorithm that has been modified to many problem areas and it is best to work for a randomly generated data points. The algorithm is composed of the following steps:
Step 1: Take k points into the space represented by the objects that are being clustered. These points represent initial group centroids.
Step 2: Allocate each item to the group that has the closest centroid.
Step 3: When all objects have been given, recalculate the positions of the k centroids.
Step 4: Repeat Steps 2 and 3 until the centroids no longer move.

However k-means is sensitive to initially randomly selected cluster centers. This can be solved by running kmeans over and over for the same or using algorithms like farthest first.

B. hybrid PSO clustering

The population-based search of the PSO algorithm reduces the effect that initial conditions has, as opposed to the K-means algorithm, the search starts from multiple positions in parallel.So the PSO algorithm performs better than the K-means algorithm in terms of quantization error In g-best PSO algorithm, we take a fitness function or optimisation function and we caluclate it over a iteration for all the elements and find the optimum value. Later to measure the performance of the clustering algorithm the cluster performance measures are used namely Random Index(RI) and Normalized Mutual Information(NMI)

IV. Data Preprocessing

Used five cancer gene expression datasets including ALLAML, Glioma, Lung, Lymphoma respectively.

Initially, we discovered all dataset in .mat format that encompassed seven types of data including header of matlab file, Platform, Created time, version, globals, values of gene expression data and labels. Upon further investigation in the values of gene expression data there were more than 2 columns so to reduce the dimensions we used Principle component analysis. So we had to convert mat files to pandas dataframes by removing headers and perform PCA to the output from conversion.

Now, the important step is to find the number of clusters for each dataset. For this two methods are used including Elbow method and Average silhouette method.

V. Experimental Results and Discussion

by performing the above methods to find the number of optimal clusters we obtained the following values for their respective datasets

TABLE I
CLUSTER

Dataset	ALLAML	Glioma	Lung	Lymphoma
No of clusters	2	4	5	9

To visualise the data properly, scatter plots are made with the data after preprocessing.

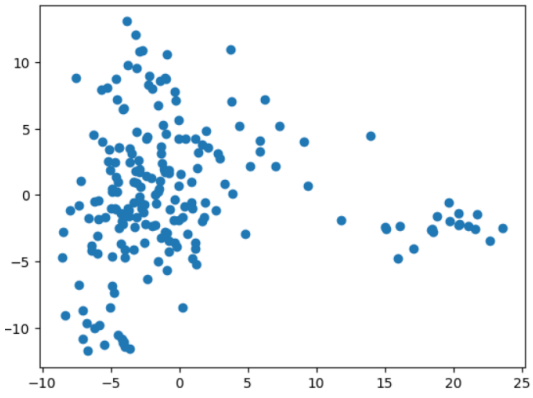

fig.1. LUNG cancer Preprocessed Data

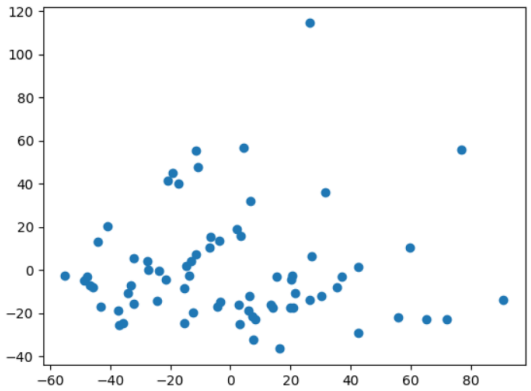

fig.2. ALLAML Preprocessed Data

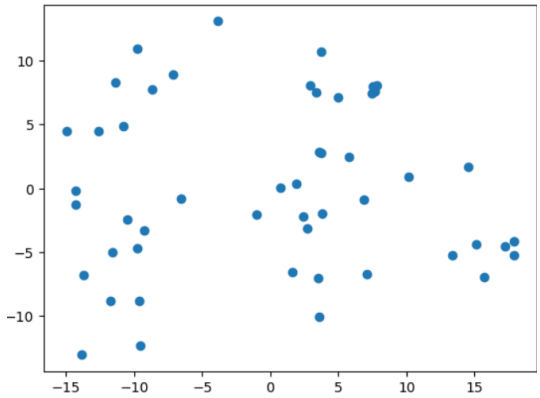

fig.3. GLIOMA Preprocessed Data

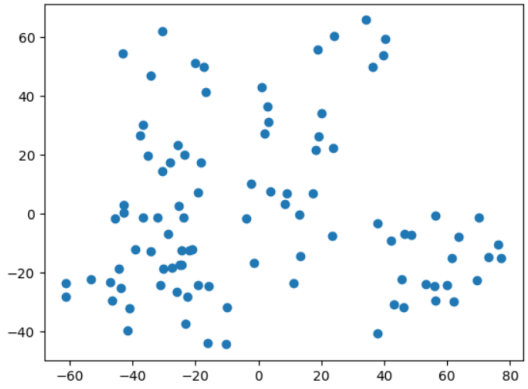

fig.4. LYMPHOMA Preprocessed Data

After using k-means algorithm on this data and calculating cluster centers using elbow and average silhouette.

DATASET	RI	NMI
ALLAML	0.60	0.15
LUNG	0.71	0.56
GLIOMA	0.73	0.50
LYMPHOMA	0.76	0.56

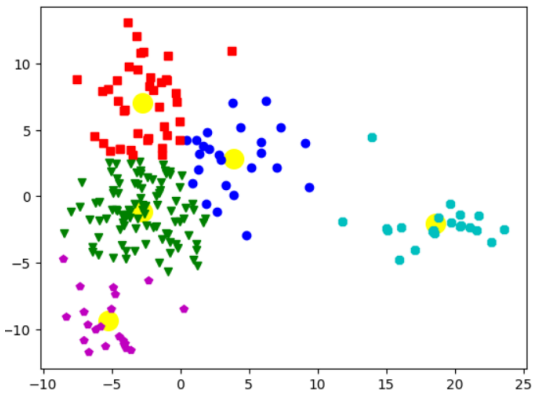

fig.5. LUNG clustered Data

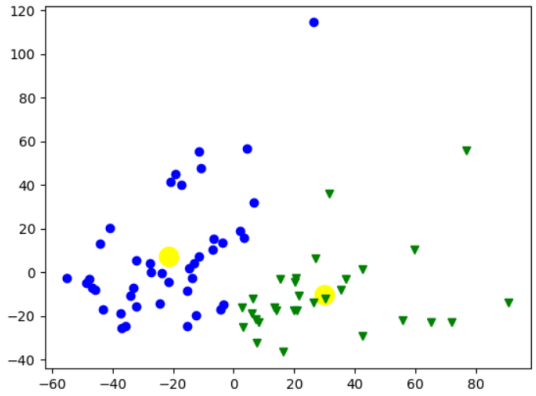

fig.6. ALLAML clustered Data

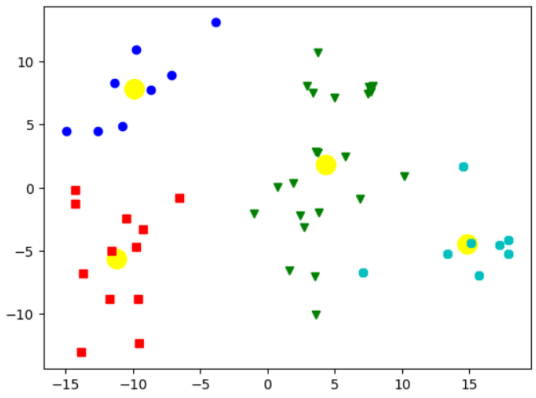

fig.7. GLIOMA clustered Data

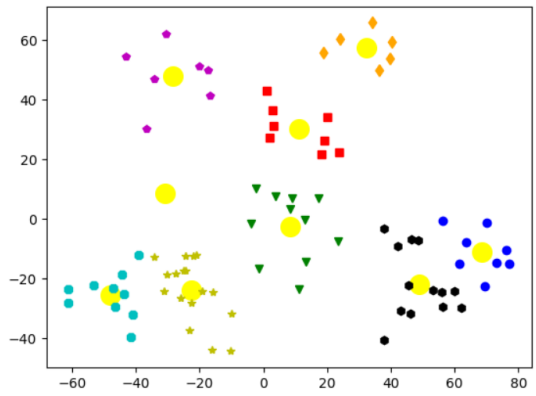

fig.8. LYMPHOMA clustered Data

VI. Conclusion

Generally, the time taken will vary from processor to processor. In this paper, we have used Hybrid PSO clustering approach along with SSE(Sum of Square error) as fitness function. The performance of the clustering algorithm was analyzed using Random Index(RI) and Normalized Mutual Information(NMI)

VII. References

.

1.Mann AK, Kaur N. Survey Paper on Clustering Techniques. IJSETR. 2013 Apr; 2(4):803–6

2.ICANN 2017: Artificial Neural Networks and Machine Learning – ICANN 2017 pp 411–419

3.Gene Expression Data Using Feature Weighted Robust Fuzzy c-Means Clustering. IEEE Trans Nanobioscience. 2022 Mar 8;PP. doi: 10.1109/TNB.2022.3157396. Epub ahead of print. PMID: 35259111.

4.Indian Journal of Science and Technology, Vol 8(15), DOI: 10.17485/ijst/2015/v8i15/73329, July 2015

VIII. Acknowledgements

We express our heartfelt gratitude to Dr. Ramalingaswamy Cheruku for his invaluable guidance during our research internship at NIT Warangal. Special thanks to the faculty, staff, fellow researchers, and participants for their support. This internship provided us with valuable insights into EEG signal classification. by the results we can conclude that the algorithm is suitable for requirement clustering of cancer related medical applications.

Dr. Ramalingaswamy Cheruku
Assistant Professor
Department of CSE
NIT Warangal