Advanced Unique Hybrid Clustering Method for effective classification of Bio-Medical Datasets

R.SatyaPrasad
Department of CS & E
Acharaya Nagarjuna University

Marri . Suneetha
Research Scholar, Department of CS & E
Acharaya Nagarjuna University

R.Mahesh
B.Tech,JNTUK

Abstract - The paper focus on a technique called PCA-UHAC-SA, which combines both the PCA techniques and unique hybrid automatic clustering algorithm for speed up the clustering analysis process by reducing the size of the tissue dataset without losing the actual original dataset. This technique is designed and implement and evaluated for better analysis of data for cancer and tumors treatments. The results are compared with PCA techniques, fuzzy c-means clustering algorithm individually with the dataset. The clustering results are obtained and displayed in base-color images and their calculating times were evaluated compared.

Keywords: Clustering, Classification, PCA-UHAC0SA.

I. INTRODUCTION

Biomedical data in clinical database has been become sensitive for health care systems, without studying the whole of the data or dataset in tissues or kidneys, the predication and evaluated of the case cannot be possible to better detect and predict and cure of damaged tissue. To Identify and predict the various parts of the tissues advanced clustering and classification algorithms has to be develop for efficient study of the tissues and its datasets organized in the tissues.

To study the tissues dataset its clinical data has been collected , the basic and most frequently technique is FTIR has become popular for investigate and quick scan on large areas of tissues and tissue section can be examined very easily and fast, compared to oral detection of cancer datasets.

Infrared imaging technique is used to capture portion of tissues in several orders of magnitude larger and, therefore, any developed technique must be capable of operating on such large datasets. This Paper focuses on the investigation of FTIR spectra data obtained by employing infrared imaging technology to analyze cancer node tissue sections. For the purpose of this Paper's study, a tissue area that incorporated a variety of lymph node tissue types was used. The most important feature of this sample area was that it contained sections of both cancerous invasion and healthy nodal tissue.

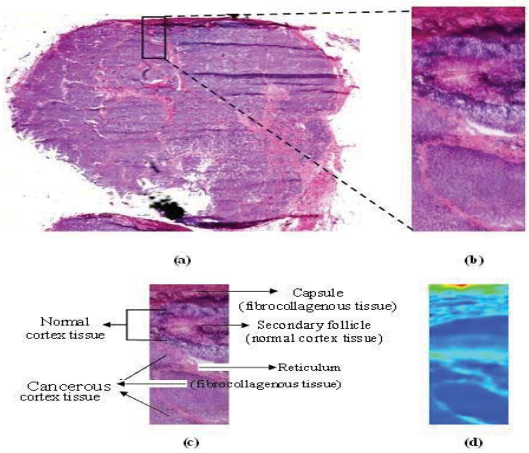


Figure 1. Photomicrograph of the (a) Large portion of Tissues section used for Infra Red analysis (b) selected area magnification (c) description of tissue positions (d) spectral image of tissue PCA can detect structure in the relationships between variables of data and can also be used to reduce the dimensionality of a dataset.

II. BACKGROUND

Babu and Murty [3] used an Evolutionary Strategy (ES) in hard and fuzzy clustering problems. In their paper, a centroid type clustering objective function was used, which enabled the approach to handle real-valued parameter optimization problems.

Kim et al [4] also recently employed PCA as a pre-processing step for cluster analysis, in a similar fashion. Lasch et al [1], using seven different species of FTIR plants spect. Zhao et al [100] investigated the use of FTIR spectroscopy to characterize a group of 20 different bacteria using the Euclidean distance as similarity measure to construct a dendrogra.

Salman et al [1] employed Ward's algorithm to look into the detection of cells infected by different variations of the herpes virus using FTIR spectroscopic methods. Zhang et al [2], a k-means clustering technique was used for the assignment of pixels in an image to identify cell and non-cell categories.

The purpose of clustering is to group the objects (spectra) so that they have the most similarity in the same cluster and objects have the most dissimilarity in the different clusters, thus, through the clustering process of different FTIR spectra, diverse types of cells can be separated. In different clustering procedures it is commonly required that the quality of the clustering schema is verified. This is achieved by a cluster validity measure. This section identified several of the most relevant cluster validity measures that have been used within the literature in order to evaluate the partition results.

III. PROPOSED WORK

The proposed word is to derive an algorithm which integrates PCA- UHAC-SA clustering analysis, where the dataset were compressed using PCA algorithms for 10 cases, PC1, PC2, PC3…PC9, with the original datasets. The new generated is then clustered using UHAC-SA methodology. The new algorithm consecutively performs 2 multivariate processes, the total calculating time is faster compared to the existing one. The new algorithm conventional states that no data loss occurs and no image quality is reduced. This image or dataset can be effectively used statistical manipulation for accurate results, which reduces computation and analysis time.

Algorithm (PCA- UHAC-SA)

1. Create a PCA variable PC1…..PC10
2. Prepare relationship between variables of data
3. Reduce the dimensionality of Spectra analysis dataset into and assign the variables
4. PCA is the main dimensional dataset and PC1 ….. to PC10 are reduced small dimensional dataset of spectra data , these 10 will reflect the original
5. Calculate the co-relation spectrum of PC's normalize it,
6. Relate the nearest spectral value, cluster the near with PC's (based on color image of spectra.
5. Apply UHAC-SA function
    - 5.1 define the string for fuzzy c-mean to generate cluster centre's from original datasets
    - 5.2 Initialize the energy function for each cluster based on index or weights(10 clusters)
    - 5.3 while Time > Time (mx)
    - For(i..k)
    - Randomly identify and alter the cluster centre with index weights of clusters
        - Calculate the new energy E_n from the clusters string
    - 5.4) If $E_n < E_c$, then accept the new string and set it as cluster current string.
    - 5.5) Else, accept the new string with a certain probability.
    - 5.6) if $E_c < E_b$, then $E_b = E_c$, and set current cluster string as the best string.
    - 5.7) End for
6) End while.
7) Return the best string as the final solution.

IV. COMPUTATIONAL TIME

Experiments and examination has been done for large spectral datasets, it is identified that the method which is used and developed takes less time in analysis i.e both fast and efficient. The technique is used repeated 10 times and the average computation times determined. All calculations were carried out on a 2.33 GHz Intel Pentium i3 PC that utilized a 3B RAM, and ran under the Windows 7 operating system. The computational time for PCA is segmented into 2 parts, they are

$\mathbf{CT_{PCA} = T_{PCA} + T_{10PCs\ IMG\ PLO}}$

T_{PCA} is time principal component, and $T_{10PCs\ IMG\ PLO}$ is the time taken to plot the 10 PCA images. The experimental results shows that T_{PCA} =62.3 seconds and $T_{10PCs\ IMG\ PLO}$ = 83.50 seconds, The total computation time is 144.9 seconds , approximately 2.5 mins.

Computation time for $CT_{UHAC\text{-}SA}$ for UHAC-SA is of two parts

$CT_{UHAC\text{-}SA} = T_{UHAC\text{-}SA} + T_{IMGPLOT(I)}$

$T_{UHAC\text{-}SA}$ is the time for UHAC-SA clustering , I represent the no of cluster (1..9) and Timg plot(i) is the time for ploting images for I clusters.

The total computation time calculated was 1759.9 seconds approx 29 minutes

Finally the computation time for PCA-UHAC-SA techniques comprises 4 parts

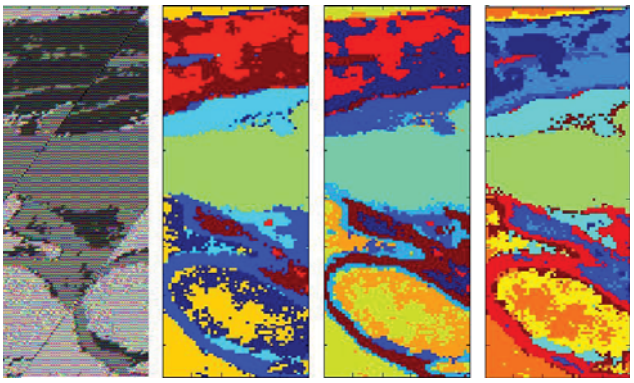
$CTPCA\text{-}UHAC\text{-}SA = T_{PCA} + T_{extract\ 10\ pcs} + T_{UHAC\text{-}SA} + T_{img\ plot(i)}$

The approximate computational time taken to cluster using this technique is 151.29 secs and approximately 2.6 mins.

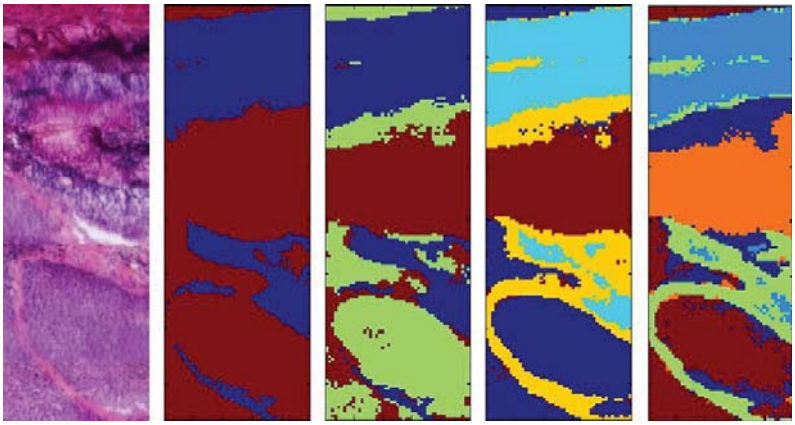
V. RESULTS AND ANALYSIS

TABLE 1. SHOWS 7 ORAL CANCER ORAL FTIR DATA SETS

Data sets	Variance Range PC1	PC2	PC3
Dataset 1	1.9458	0.9141	0.3575
Data set 2	2.9569	1.5562	0.9795
Data set 3	1.999	0.9859	0.96
Data set 4	2.7702	1.9969	0.6112
Data set 5	1.7224	0.9317	0.8496
Data set 6	1.9023	1.0902	0.5342
Data set 7	1.975	1.79	0.9741



(a) 2clusters (b) 3clusters (c) 4clusters (d) 5clusters



(f) 6clusters (j) 7clusters (h) 8clusters (i) 9clusters

Figure 2. IR imaging of lymph node tissue section LNII5 by (PCA- UHAC-SA (a) H&E stained image of LNII5 (b)−(i false colour weighted clustering results, the number of clusters were from 2 – 9 respectively

Table 2 Summary of PCA- UHAC-SA computation time.

Number of Clusters	Computation time for **PCA- UHAC-SA** technique		
	fuzzy c-means clustering (sec)	Image plot (sec)	Total
2	0.67	0.36	1.03
3	4.15	0.37	4.52
4	3.34	0.34	3.68
5	4.31	0.39	4.7
6	9.83	0.42	10.25
7	7.9	0.39	8.29
8	12.63	0.39	13.02
Subtotal	42.83	2.66	45.49
^{T}PCA	-	-	60.3
^{T}extract10PCs	-	-	0.01
Total	-	-	151.29

TABLE 3. Comparison with the existing Techniques

Techniques	**Computation times(mins)**
PCA	**2.5**
FUZZY C-means	**2.8**
PCA- UHAC-SA	**2.48**

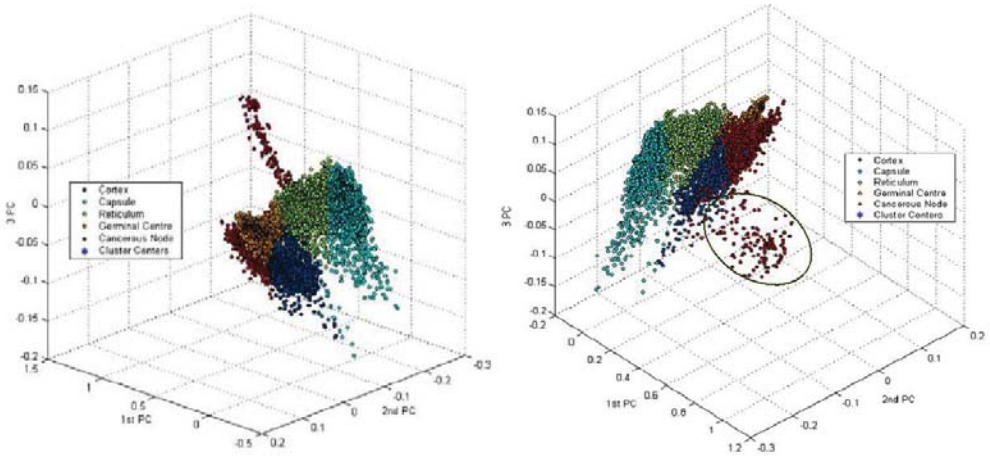


Figure: 3 LNII5 tissue section spectra plot in three dimensional PCs space (a) original plot with 5 clusters (b) rotated plot of picture (a).

At the outset , The Combined **PCA- UHAC-SA** analysis achieve higher degree compared to the three, but it is similar to the Fuzzy c-mean, the results show that it has improved evaluation speed without loss of original data. This also provides high quality clustered data from large datasets for evaluation of cancer tissues.

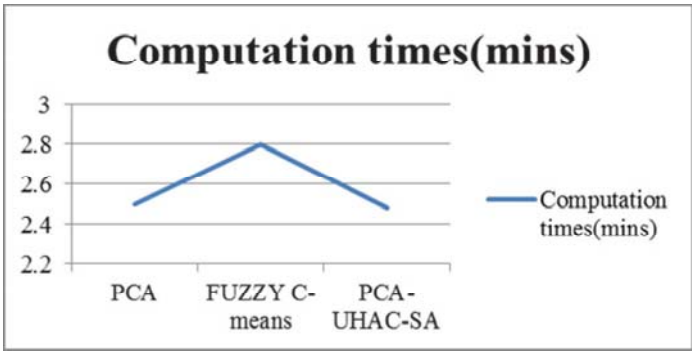


Figure 4. : Computation Time comparison for Clustering Techniques

Figure 4 shows the computation time comparative report of all three clustering techniques based on the 7 datasets used of FITR datasets on oral cancer. The graph shown above , give the time execution of clustering technique, It requires less time to cluster the dataset with high rate of accuracy.

VI. CONCLUSION

Tissue data was collected from IR image , Due to large size of data, information processing is very difficult, In my paper , I present **PCA- UHAC-SA clustering** technique which combines both in reducing the data which other altering the futures of the datasets.

The dataset reduced is compared with the existing Fuzzy C-mean and PCA. Based on the comparison **PCA- UHAC-SA** is 10 times faster and can reduce the size of datasets without loss of information that is available in original datasets.

REFERENCES

[1] Zhang, L., Small, G. W., Haka, A. S., Kidder, L. H., and Lewis, E. N., 2003, "Classification of Fourier Transform Infrared Microscopic Imaging Data of Human Breast Cells by Cluster Analysis and Artificial Neural Networks", Applied Spectroscopy, vol. 57, no. 1, pp. 14-22.

[2] Babu, G. P. and Murty, M. N., 1994, "Clustering with Evolution Strategies", Pattern Recognition, vol. 27, no. 2, pp. 321-329.

[3] Kim, S. W., Ban, S. H., Chung, H., Cho, S., Chung, H. J., Choi, P. S., Yoo, O. J., and Liu, J. R., 2004, "Taxonomic Discrimination of Flowering Plants by Multivariate Analysis of Fourier Transform Infrared Spectroscopy Data", Plant Cell Reports, vol. 23, no. 4, pp. 246-250.

[4] Zhao, H., Kassama, Y., Young, M., Kell, D. B., and Goodacre, R., 2004, "Differentiation of Micromonospora Isolates from a Coastal Sediment in Wales on the Basis of Fourier Transform Infrared Spectroscopy, 16S rRNA Sequence Analysis, and the Amplified Fragment Length Polymorphism Technique", Applied and Environmental Microbiology, vol. 70, no. 11, pp. 6619-6627.