ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

Comparison of Hybrid Unique clustering techniques

Marri. Suneetha[1], R.Satya Prasad[2] , R.Mahesh[3]

Research Scholar, Department of CSE, ANU, Andhra Pradesh, India

Department of CS & E, Acharaya Nagarjuna University, Andhra Pradesh, India

B.Tech Student, JNTU University, Andhra Pradesh, India

ABSTRACT: Clustering is the most important technique in data mining. In data mining, clustering is very useful to discover various similar patterns in the underlying data. The implementation of clustering algorithms are distance metric based similarity measure in order to partition the database such that data points in the same partition are more similar than points in different partitions. In this paper, the comparison of six clustering algorithms namely hierarchical cluster analysis, k-means and fuzzy c-means algorithms, VFC-SA, UHAC-SA, PCA-UHAC-SA in infrared analysis on seven oral cancerous FTIR datasets. The results show the performance of all the proposed algorithms and its implementation.

KEYWORDS: hierarchical cluster analysis (HCA), k-means and fuzzy c-means algorithms, VFC-SA, UHAC-SA, PCA-UHAC-SA.

I.INTRODUCTION

Cancer has become a major adversary to human health, and the development and enhancement of techniques for use in its diagnosis and treatment has increasingly become a focus of worldwide research. Fourier Transform Infrared (FTIR) spectroscopy is a powerful tool for determining the biochemical composition within a biological system. This capability to provide an insight into the biochemical changes that occur within cells has led, in recent years, to FTIR spectroscopy being investigated in the study of various biomedical conditions.

In a real medical diagnostic application, for a previously unseen tissue sample, the number of different types of cells is normally not known in advance. Based on this fact, a clustering technique which can automatically obtain the appropriate number of tissue types is required. There have been many clustering methods have been developed in attempt to automatically determine the optimal number of clusters. Recently, Bandyopadhyay proposed a Variable String Length Simulated Annealing (VFC-SA) algorithm [1], which applied a simulated annealing algorithm to the fuzzy c-means clustering technique and used a cluster validity index measure as the energy function. This has the advantage that, by using simulated annealing, the algorithm can escape local optima and, therefore, may be able to find the globally optimal solution(s). The D-VC index was used as the cluster validity index to evaluate the quality of the solutions; the author stated that this is because it has been shown to be able to detect the correct number of clusters in several experiments [2].

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.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

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.

In this paper, the comparison between hierarchical cluster analysis, k-means and fuzzy c-means algorithms, VFC-SA, UHAC-SA, PCA-UHAC-SA are described in detail. The experiments in this paper were performed on the same seven FTIR spectra datasets containing oral cancer cells in order to evaluate the performance of the VFC-SA and UHAC-SA clustering algorithms in comparison with the original fuzzy c-means algorithm and PCA-UHAC-SA. The study and description of Oral Cancer Datasets are given in [3].

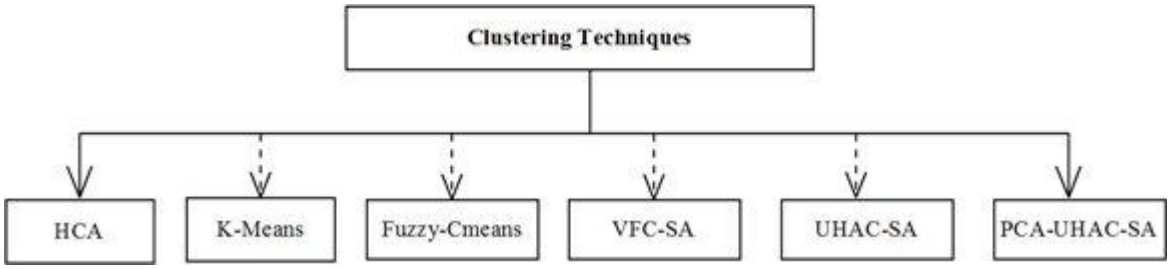

Figure-1, Clustering Techniques

2. Oral Cancer Datasets

Our comparison is totally based on the FTIR spectra datasets containing oral cancer cells. The main objective of the paper to provide the comparison of all the algorithms and identify the unique hybrid clustering algorithm which improves the performance of the clustering algorithms [1] [15].

II.HCA ALGORITHM OF CLUSTERING

1. Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain [4].
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*).

K-means: The Squared Euclidean distance was used to compute the distance between each data point to its centroid; Maximum number of iterations was 100.

Algorithmic steps for k-means clustering

Let X = {x1,x2,x3,……..,xn} is data points set and V = {v1,v2,…….,vc} is centers set.

1. Randomly select ‘c’ cluster centers.
2. Calculate the distance between each data point and cluster centers.
3. Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..
4. Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, ‘ci’ represents the number of data points in ith cluster.

5. Recalculate the distance between each data point and new obtained cluster centers.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

6. If no data point was reassigned then stop, otherwise repeat from step 3).

Fuzzy c-means: Fuzziness index is equal to 2; maximum number of iterations was 100; minimal amount of improvement was 10 to 5.

Similarly to k-means, the squared Euclidean distance was also used to calculate the distances between data points to centroids.

Algorithmic steps for Fuzzy c-means clustering

Let X = {x1, x2, x3 ..., xn} be the set of data points and V = {v1, v2, v3 ..., vc} be the set of centers.

1. Randomly select 'c' cluster centers.
2. Calculate the fuzzy membership 'μij' using:

$$\mu_{ij} = 1 / \sum_{k=1}^{c} (d_{ij} / d_{ik})^{(2/m-1)}$$

3. Compute the fuzzy centers 'vj' using:

$$v_j = (\sum_{i=1}^{n} (\mu_{ij})^m x_i) / (\sum_{i=1}^{n} (\mu_{ij})^m), \forall j = 1, 2, ...c$$

4. Repeat step 2) and 3) until the minimum 'J' value is achieved or ||U(k+1) - U(k)|| < β.

where,

'k' is the iteration step.
'β' is the termination criterion between [0, 1].
'U' = (μij)n*c is the fuzzy membership matrix.
'J' is the objective function.

The implementation of these algorithm were performed using Matlab (version 6.5.0).

Table: 1 Distribution of the different tissue types identified clinically and as obtained by the various clustering techniques

Dataset Names	Tissue Types	Hierarchical Clustering				K - means			Fuzzy c-means	
		Single	Average	Complete	Ward					
Dataset 1	Tumour	0	0	0	0	0			0	
	Stroma	0	0	0	0	0			0	
Dataset 2	Tumour	7	0	0	0	0			0	
	Stroma	0	1	1	1	1			1	
Dataset 3	Tumour	0	0	0	0	0	0	0	0	
	Stroma	4	4	5	3	5	2	4	4	
Dataset 4	Tumour	7	7	3	3	3	7	3	3	7
	Stroma	5	5	3	3	4	5	2	4	5
	Early keratinisation	0	0	0	0	0	0	0	0	0
Dataset 5	Tumour	12	0	0	0	0	0		0	
	Stroma	1	0	0	0	4	1		4	
Dataset 6	Tumour	0	0	0	0	0			0	
	Stroma	0	0	0	0	0			0	
Dataset 7	Tumour	7	0	0	0	0			0	
	Stroma	0	4	4	6	4			2	
	Necrotic	1	0	0	0	0			1	

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

In order to further investigate the performance of these different clustering methods, the average number of disagreements for all datasets were calculated, as shown in Table 4.4. It can be seen that the hierarchical clustering single linkage method has the worst performance, the average linkage performance is better to single linkage, while the complete linkage and ward methods performed better overall. K-means and fuzzy c-means have fairly good performance, approximately linear with *n*. Hence, compared with hierarchical clustering, these techniques will be far less time-consuming on large datasets. Hence, the overall conclusion is that fuzzy c-means is the most suitable clustering method in this context.

Table 2 Average number of disagreements obtained in the three clustering methods

	Hierarchical Clustering				K - means	Fuzzy c-means
	Single	Average	Complete	Ward		
Average (SD) Number of disagreements per run	44	21	16	16	18.8±5.8	19.5±1.6
Average (SD) Number of disagreements per run per dataset	6.3	3.0	2.3	2.3	2.7±0.8	2.8±0.2

III.VFC-SA CLUSTERING ALGORITHM

In this algorithm, a variable number of cluster centres were encoded using a variable length string to which simulated annealing was applied. The process of altering the current cluster centres comprised three functions. They are: perturbing an existing centre (*Perturb Centre*), splitting an existing centre (*Split Centre*) and deleting an existing centre (*Delete Centre*). At each iteration, one of the three functions was randomly chosen. When splitting or deleting a centre, the cluster sizes were used to select a centre. The size, C_j , of a cluster, j , can be expressed by:

$$|\mathbf{Cj}| = \sum_{i=1}^{n} \propto_{ij}, \quad j = 1, \dots \dots c \qquad (1)$$

where 'c' is the number of clusters. The three functions are described below.

1) Set parameters T_{max} , T_{min} , c, k, r.
2) Initialised the string by randomly choosing c data points from the dataset to be cluster centres.
3) Compute the corresponding membership values using Equation **(2)**
4) Calculate the initial energy E_c using V_{xb} index from Equation **(3)**
5) Set the current temperature $T = T_{max}$.
6) While $T \geq T_{\min}$
   6.1) For i □ 1to$k$
      6.1.1) randomly alter a current centre in the string
      6.1.2) Compute the corresponding membership values using Equation **(2)**.
      6.1.3) Compute the corresponding centres with the equation
      6.1.4) Calculate the new energy E_n from the new string.
      6.1.5) If $E_n < E_C$, then accept the new string and set it as current string.
      6.1.6) Else accept the new string with a certain probability.
   6.2) End for
   6.3) T □ rT .
7) End while.
8) Return the current string as the final solution.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

IV.UHAC-SA CLUSTERING ALGORITHM

When the original VFC-SA algorithm was implemented on a wider set of test cases than used by the original authors [3], it was found to suffer from several difficulties.

The first extension is in the initialization of the string. Instead of the original initialization in which random data points were chosen as initial cluster centres, the fuzzy c-means clustering algorithm was applied using the random integer c $[c_{\min}, c_{\max}]$ as the number of clusters.

The second extension is in *Perturb Centre*. The method of choosing a centre in the VFC-SA algorithm is to randomly select a centre from the current string. However, this means that even a 'good' centre can be altered.

The third extension is in *Split Centre*. If the boundary between the biggest cluster and the other clusters is not obvious (not very marked), then the approach that original authors use is to choose a reference point with a membership degree that is less than but closest to 0.5.

Solutions for the seven FTIR datasets were generated by using the fuzzy c-means, VFC-SA and UHAC-SA algorithms. Each method was allowed 10 runs on each dataset. As mentioned at the beginning of this section, the number of clusters was predetermined for fuzzy c-means through clinical analysis. The outputs of fuzzy c-means (centres and membership degrees) then used to compute the corresponding V_{XB} index value. VFC-SA and UHAC-SA automatically found the number of clusters by choosing the solution with the smallest V_{XB} index value. Table 5.1 shows the average V_{XB} index values obtained after ten runs of each algorithm (best average is in bold).

Table 3 Average of the V_{XB} index values obtained when using the fuzzy c-means, VFC-SA and UHAC-SA algorithms.

Dataset	Average V_{XB} Index Value		
	Fuzzy C-Means	**VFC-SA**	**UHAC-SA**
1	0.048036	0.047837	0.047729
2	0.078896	0.078880	0.078076
3	0.291699	0.282852	0.077935
4	0.416011	0.046125	0.046108
5	0.295937	0.251705	0.212153
6	0.071460	0.070533	0.070512
7	0.140328	0.149508	0.135858

From Table 3, it can be seen that in all of these seven datasets, the average V_{XB} values of the solutions found by UHAC-SA are smaller than both VFC-SA and fuzzy c-means. This means that the clusters obtained by UHAC-SA have, on average, better V_{XB} index values than the other two approaches. Put another way, it may also indicate that UHAC-SA is able to escape sub-optimal solutions better than the other two methods.

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.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

7. Apply UHAC-SA function
   7.1. define the string for fuzzy c-mean to generate cluster centre's from original datasets
   7.2. Initialize the energy function for each cluster based on index or weights(10 clusters)
   7.3. 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
   7.4. If En <Ec , then accept the new string and set it as cluster current string.
   7.5. Else, accept the new string with a certain probability.
   7.6. if Ec < Eb , then Eb = Ec , and set current cluster string as the best string.
   7.7. End for
8. End while.

Return the best string as the final solution.

Table 4. Comparison with the existing Techniques

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

At the outset, the combined **PCA- UHAC-SA** analysis achieve higher degree compared to the other 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.

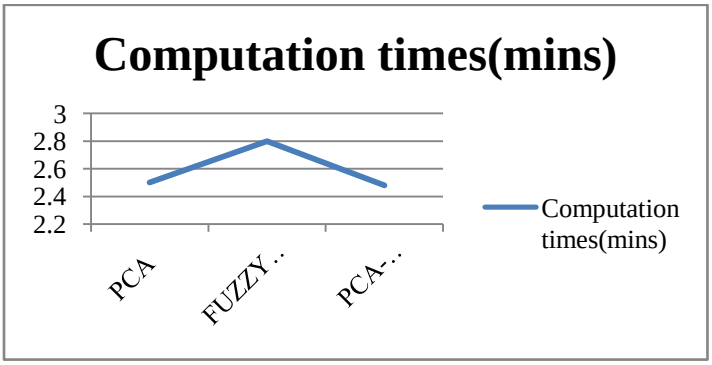

Figure 2 Computation Time comparison for Clustering Techniques

Figure 6.4 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.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

V.CONCLUSION

The comparison of all the clustering techniques based on the 7 datasets used of FTIR datasets on oral cancer the implementation done on all the algorithms with the given dataset. The objective of all the algorithms is to provide HCA with some features formed the clusters; k-means is one of the clustering techniques which provide the better similar clusters in cancer dataset, performance of these fuzzy c-means different clustering methods, the average number of disagreements for all datasets were calculated fuzzy c-means have fairly good performance. a new UHAC-SA method has been proposed which has been extended from the original VFC-SA algorithm in four ways. The newly proposed algorithm's performance has been evaluated on seven oral cancer FTIR spectra data and compared to clinical analysis, the standard fuzzy c-means and the original VFC-SA. The XB validity index was used as the evaluation method to measure the quality of the clusters produced. PCA- UHAC-SA clustering technique which combines both in reducing the data which other altering the futures of the datasets.

REFERENCES

[1] Allibone, R., Chalmers, J. M., Chesters, M. A., Fisher, S., Hitchcock, A., Pearson, M., Rutten, F. J. M., Symonds, I., and Tobin, M., 2002, "FT-IR microscopy of oral and cervical tissue samples", Derby City General Hospital, Internal Report.
[2] Lasch, P., Haensch, W., Naumann, D., and Diem, M, 2004, "Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis", Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, vol. 1688, no. 2, pp. 176-186.
[3] Berkhin, P., 2002, "Survey of Clustering Data Mining Techniques", San Jose, CA,USA, Accrue Software.
[4] A Novel Methods of Investigation on Bio-Medical Cancer Tissues using Advanced Clustering Techniques, R.SatyaPrasad, Marri. Suneetha, R.Mahesh, International Journal of Computer Applications (0975 – 8887) Volume 129 – No.12, November2015.
[5] 1999, "Characteristics of Methods for Clustering Observations",http://www.id.unizh.ch/software/unix/statmath/sas/sasdoc/stat/chap8/sect4.htm, SAS/STAT User's guide onlineDoc,Version 8, SAS Institute Inc.,Cary,NC,USA.
[6] Richter, T., Steiner, G., Abu-Id, M., Salzer, R., Bergmann, R., Rodig, H., and Johannsen, B., 2002, "Identification of Tumor Tissue by FTIR Spectroscopy in Combination with Positron Emission Tomography", Vibrational Spectroscopy, vol. 28, pp. 103-110.
[7] Lasch, P., Haensch, W., Naumann, D., and Diem, M, 2004, "Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis", Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, vol. 1688, no. 2, pp. 176-186.
[8] R. Satya Prasad, Marri. Suneetha, R. Mahesh , A Novel Methods of Investigation on Bio-Medical Cancer Tissues using Advanced Clustering Techniques, *International Journal of Computer Applications (0975 – 8887) Volume 129 – No.12, November2015.*
[9] Bezdek, J., 1998, *Pattern Recognition in Handbook of Fuzzy Computation*, IOP Publishing Ltd. Boston, NY.
[10] Jolliffe, I.T., 1986, *Principal Component Analysis*, Springer-Verlag. New York.
[11] Bandyopadhyay, S., 2003, "Simulated Annealing for Fuzzy Clustering: Variable Representation, Evolution of the Number of Clusters and remote Sensing Applications", unpublished, private communication.
[12] Pal, N. R. and Bezdek, J., 1995, "On Cluster Validity for the Fuzzy C-Means Model", *IEEE Trans.Fuzzy System.*, vol. 3, pp. 370-379.
[13] Rayward-Smith, V.J., Osman, I.H., Reeves, C.R., and Smith, G.D., 1996, *Modern Heuristic Search Methods*, John Wiley & Sons.
[14] Conover, W.J., 1999, *Practical Nonparametric Statistics*, John Wiley & Sons.
[15] Causton, D.R., 1987, *A Biologist's Advanced mathematics*, Allen & Unwin. London.
[16] A Unique Hybrid Automatic Clustering Method for Dynamic Determining the Number of Cluster and Classify the Clusters- UHAC-SA, R.SatyaPrasad, Marri. Suneetha, R.Mahesh, Vol. 4, Issue 3, March 2016.
[17] 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.
[18] Babu, G. P. and Murty, M. N., 1994, "Clustering with Evolution Strategies", Pattern Recognition, vol. 27, no. 2, pp. 321-329.
[19] 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.
[20] 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.