A Novel Methods of Investigation on Bio-Medical Cancer Tissues using Advanced Clustering Techniques

R. Satya Prasad
Department of CS & E,
Acharaya Nagarjuna University

Marri. Suneetha
Research Scholar,
Department of CS & E

R. Mahesh
B.Tech,JNTUK

ABSTRACT

The paper investigates the performance of three often used clustering techniques, namely hierarchical cluster analysis, k-means and fuzzy c-means algorithms, in infrared analysis on seven oral cancerous FTIR datasets. The diagnostic results from clinical study are considered as the ‘gold standard’ in this paper the proposed system evaluates the clustering results from these three techniques. Corresponding experiments were carried out and the results showed that fuzzy c-means is the most suitable clustering method in this context.

Keywords

clustering, FTIR, bio-medical.

1. 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 order to analyze the FTIR spectroscopic data from tissue samples, multivariate clustering techniques have often been used to separate sets of unlabelled infrared spectral data into different clusters based on their characteristics. The purpose of clustering is to group the spectral data such that the data in the same clusters are as similar as possible and data within different clusters are as dissimilar as possible. Hence, different types of cells can be separated within biological tissue. Among existing clustering techniques, it identify the best clustering method.

At first analysis of sets of FTIR spectra taken from oral cancer tissue samples [26]. In general, the

experiments analyzed the tissue samples in two parallel processes. In the first process, the samples were scanned by FTIR spectroscopy and pre-processing procedure were applied to the output of spectra from IR spectrometry. Furthermore, a set of extra various pre-processing techniques, such as mean-centering, variance scaling and first derivative were also performed on the FTIR spectra empirically for the specific multivariate analysis in order to utilise classification of different tissue types. HCA (average linkage) was mainly used to classify the spectral data from different types of tissue area and PCA was used to distinguish these data by visual inspection. In the second process, the samples were stained with a chemical solution and then examined through conventional cytology to group the samples into different functional groups. The outcomes from these two processes were then compared. The clustering results showed that accurate clustering could only be achieved by manually applying extra pre-processing techniques that varied according to the particular sample characteristics and clustering algorithms. However, the pre-processing procedures needed extra time, software tools and significant human expertise. If a clustering technique could be developed which could obtain clustering results as good or even better than conventional clinical analysis without the necessity for pre-processing procedures, it would make the diagnosis more efficient and enable automation.

2. ORAL CANCER DATASETS STUDY AND DESCRIPTION

In the oral cancer FTIR spectra, there are a total of seven datasets taken from three different patients. The spectral range in this study was limited to a 900−1800cm-1 interval. Figure 1 (a) shows a 4× magnification visual image from one of Hematoxylin and Eosin stained oral tissue sections. There are two types of cells (stroma and tumour) in this section with their regions are clearly identifiable by their light and dark coloured stains respectively. Figure 1 (b) shows a 32× magnified visual image from a portion of a parallel, unstained section; the superimposed dashed white line separates the visually different morphologies. Five single point spectra were recorded from each of the three distinct regions. The locations of these are marked by “+” on Figure 1 (b) and numbered as 1−5 for the upper tumour region, 6−10 for the central stroma layer, and 11−15 for the lower tumour region The fifteen FTIR transmission spectra from these positions are recorded as dataset 1, and the corresponding FTIR spectra (without extra pre-processing) are shown in Figure .2.

.

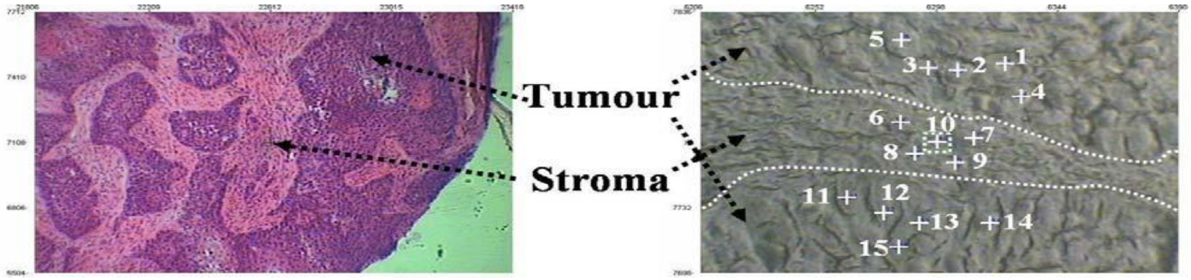


Figure 1 Tissue sample from Dataset 1; (a) 4× stained picture; (b) 32× unstained picture

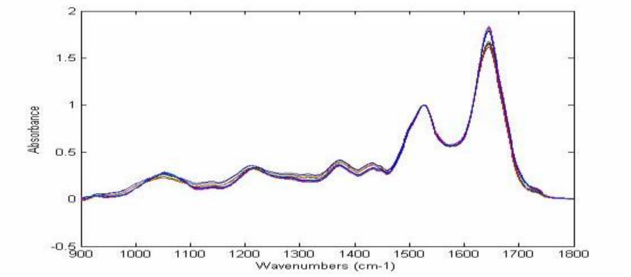


Figure 2 FITR spectra from Dataset 1.

Figure.3 shows a 32× magnified visual image of dataset 2, unstained section; the superimposed dashed white line separates the visually different morphologies. Ten single point spectra numbered as 16−25 on the right hand side for the tumour region, and rest of eight spectra numbered 26−33 on the left hand side for the stroma region.

Figure.4 shows a 32× magnified visual image of dataset 3, unstained section. There are also two types of cells (stroma and tumour) in this section with their regions. Four spectra numbered as 34−37 for the left tumour region, three spectra numbered as 38−40 from the central stroma layer, and rest of four spectra numbered as 41−44 from the right tumour region.

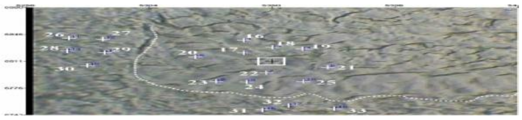


Figure 3 32× unstained picture from tissuesample Dataset 2

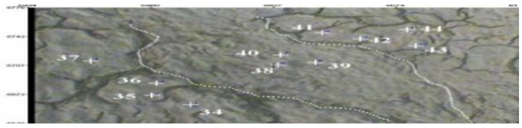


Figure 4 32× unstained picture from tissue sample Dataset 3.

Figure 5 shows a white light image of three types of tissue sample from dataset 4 and different morphologies can be visualised in the picture. The corresponding spectra numbers are also shown below (The distinct grey-scale contrast between the left half and right half of the image is artificial. It is a consequence of the image being a composite of two independent pictures corresponding to each half). It may be noticed that the boundary between stroma and early keratinisation follows a meandering way through area numbers 88, 72, 56 and 55; and in a similar manner, the boundary between the marked tumour and stroma region does not follow a vertical line as indicated, but rather appears to meander somewhere through the area contained within the area numbers 50−52, 65−67 and 80−82. A closer histopathological inspection highlighted that there had been invasion of the stroma region by tumour within the vicinity of the boundary between the two layers. At this stage of the study, we are only concerned with ascertaining spectral characteristic of essentially distinct classes of tissue cells, rather than gradation processes or mixed types [1]. Therefore those within the two boundary regions and invasion area are excluded. These spectra number include: 46, 50, 51, 55, 56, 65, 66, 71, 72, and 81−88. That is, the number of spectra was reduced from the original 48 to 31. Subsequently, the corresponding spectral points were renumbered sequentially from 45 to 75. The three different categories of tissue types in the new spectral numbering are as follows: tumour: 45−48, 56−59, 68−71.

Stroma: 49−51, 60−63.

Early keratinisation: 52−55, 64−67, 72−75.

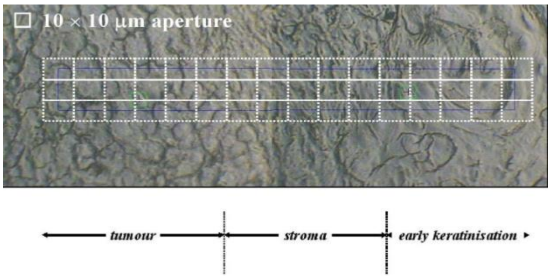
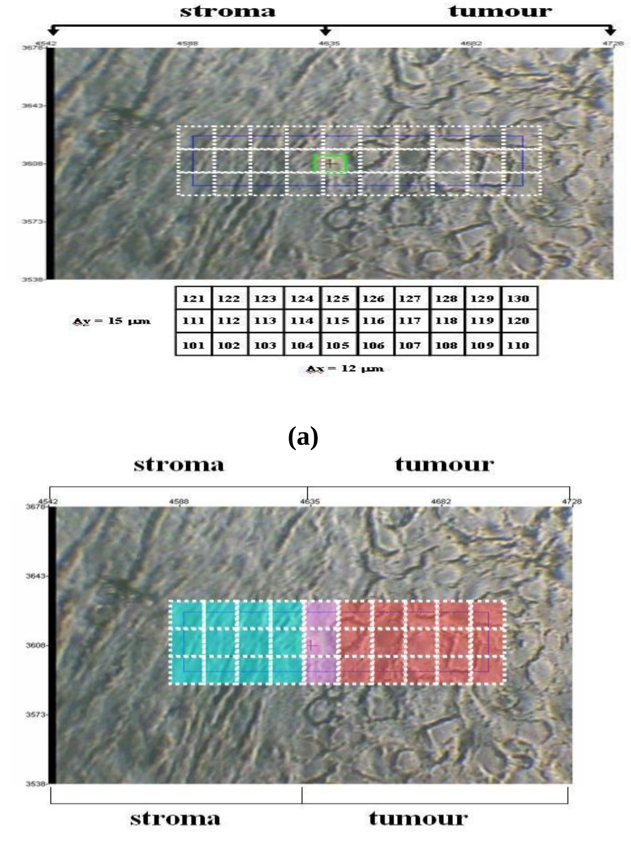


Figure 5 White light image of tissue sample Dataset 4.

78	**79**	**80**	**81**	**82**	**83**	**84**	**85**	**86**	**87**	**88**	**89**	**90**	**91**	**92**
62	**63**	**64**	**65**	**66**	**67**	**68**	**69**	**70**	**71**	**72**	**73**	**74**	**75**	**76**
46	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**

Figure 6 presents a tissue sample from dataset Thirty spectra were recorded in each grid on the white light image and their corresponding spectra numbers are also displayed in Figure 6 (a). Figure 6(b) shows the same tissue area "spectroscopic-staining" image according well with that from conventional histopathology H&E staining. Two types of tissue cells (stroma and tumour) exists in this section, however in the boundary region coloured as purple in Figure 6(b) were closer to tumour than stroma through the analysis.

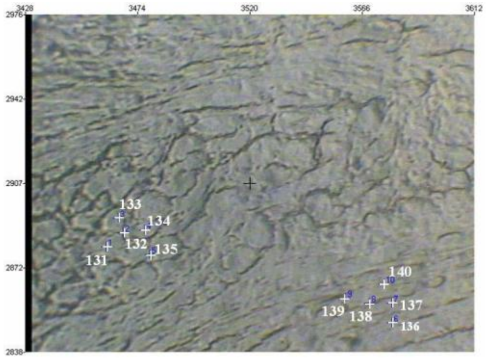


(a)

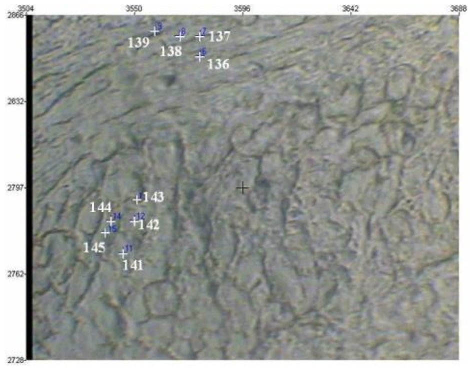
(b)

Figure 6 Tissue section from dataset 5 (a) white light image (b) spectroscopic-staining image

Figure 7 displays a tissue section and fifteen spectra, taken in two images. Three visually different areas numbered as 131−135, 136−140 and 141−145 are associated with characteristic of tumour, stroma and tumour respectively.



(a)



(b)

Figure 7 White image of tissue sample for dataset 6 (a) part 1 (b) part 2.

3. EXPERIMENTS ON ORAL CANCER DATASETS

3.1 Methodology

In this Paper, three data clustering techniques that have often been used in FTIR spectroscopy analysis, namely hierarchical cluster analysis (HCA), k-means and fuzzy c-means clustering, are used to classify the seven oral cancer FTIR spectra datasets introduced above. These had been obtained through conventional cytology [1], and no further extra pre-processing was applied (only basic pre-processing). In hierarchical clustering, four different types of linkage methods, namely "single", "average", "complete" and "ward" were conducted individually. Due to the k-means and fuzzy c-means algorithms being sensitive to the initial states, each method was run ten times. The parameters setting for these three clustering algorithms that were used are as follows:

Figure 8 shows a set of five white light images taken from an oral tissue section from a third patient. Histopatholgical examination showed that this is a complex region containing stroma, tumour and necrotic tissue. A linear scan consisting of consecutive spectral points was recorded. Similar to dataset 4, some spectra which lie in a boundary between cell types or have spectral characteristic which are not clear were eliminated from the original recorded points, leaving 42 spectra. After readjustment of the numbering, the spectra are distributed as follows:

Tumour: 201−210, 225−235.

Stroma: 211−224.

Necrotic: 236−242.

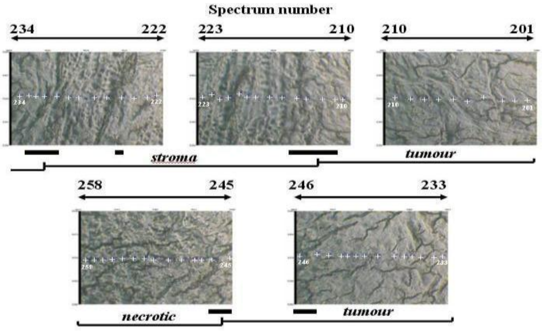


Figure 8 White image of tissue sample for Dataset 7

HCA: The Euclidean distance was used to calculate the distance between different data points.

4. 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.
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.

4.1 Algorithmic steps for k-means clustering

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

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.

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 10-5.

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, release 13.0.1).

4.2 Implementation

The distribution of the numbers of different types of tissue identified clinically and as obtained by the three clustering techniques are displayed in Table 1. As mentioned previously, clustering is an unsupervised process; this means that the results of the clustering are simply to group the data into two or more unlabelled categories. In the results presented below, the clusters were mapped to the actual known classifications in such a way as to minimise the number of disagreements from clinical studies in each case. The results are presented in comparison with a previous study on the same data where the data was pre-processed empirically before a diagnosis analysis. In this study, all three clustering analyses were performed using MATLAB (version 6.5.0, release 13.0.1).

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

Datasets 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	

After running each clustering technique ten times, it can be seen that the k-means and fuzzy c-means algorithms obtained more than one clustering result in some datasets. This is because different initialisation may lead to different partitions for both of these algorithms. From Tables 1 and 2, k-means has more variations (3 out of 7 datasets) than fuzzy c-means (1 out of 7 datasets), and their corresponding frequency (out of 10 runs) is shown in Table 3

Table 3 Clustering variations for k-means andfuzzy c-means within three datasets.

Datasets Names	***K-means***			***Fuzzy c-means***	
Dataset 3	2/10	3/10	5/10	-	
Dataset 4	3/10	3/10	4/10	9/10	1/10
Dataset 5	5/10	5/10		-	

In order to further investigate the performance of the different clustering methods, the average number of disagreements for all datasets was calculated, as shown in Table 4. It can be seen that the hierarchical clustering single linkage method has the worst performance, the average linkage performance is better than single linkage, while the complete linkage and ward methods perform the best overall, However, hierarchical clustering techniques are computationally expensive (proportional to n2, where n is the number of spectral data), therefore, they are not suitable for very large datasets [2]. K-means and fuzzy c-means have fairly good performance, and for both the computational effort is approximation linearly with n. Hence, compared with hierarchical clustering, these techniques will be far less time-consuming on large datasets [2]. Moreover, although k-means has a slightly better performance than fuzzy c-means (slightly fewer disagreements, on average), it can be seen from the standard deviations in Table 4 that k-means exhibits more variation in its results than fuzzy c-means. Hence, the overall conclusion is that fuzzy c-means is the most suitable clustering method in this context.

Table 4 Average number of disagreements obtained in the three clustering methods.

	Hierarchical clustering				**K-means**	**Fuzzy c-means**
	Single	*Average*	*Complete*	*Ward*		
Average (S.D.) Number of	44	21	16	16	18.8±5.8	19.5±1.6

Disagreements per Run						
Average (S.D.) Number of Disagreements per Run per Dataset	6.3	3.0	2.3	2.3	2.7±0.8	2.8±0.2

5. CONCLUSION

In this paper, the proposed system investigated seven sections of tissue samples containing oral cancer cells using two comparative parallel processes: that is, histological analysis and FTIR spectroscopy with the subsequent application of multivariate analysis .Prior to the multivariate analysis, all spectral data had to be empirically pre-processed. It was found that accurate clustering could only be achieved by manually applying extra pre-processing techniques that varied according to the particular sample characteristics. Furthermore, these pre-processing methods required additional time, software tools and significant human expertise to perform. In this paper, three commonly used clustering techniques in FTIR spectroscopic data analysis, namely, HCA, k-means and fuzzy c-means were applied to the same seven spectral datasets as Chalmers et al reported but without any extra pre-processing procedure. Single, average, complete linkage and Ward’s method were employed in the HCA clustering techniques.

The experimental results showed that the single linkage method obtained the worst clustering results, average linkage method’s performance was better than single linkage but, overall, complete linkage and Ward’s method obtained the best of the solutions. However, one of major drawback for HCA clustering algorithm is high computation expense. Therefore, for very large datasets (which normally appear in practical FTIR spectral analysis), this method may not be suitable. On the other hand, the k-means and fuzzy c-means algorithms performances also achieved the good performance. In addition, they require less computational resources in comparison with the HCA method. However, from the clustering results it can be seen that k-means clustering algorithm generated less consistent clustering results than fuzzy c-means. Overall, it may be suggested that fuzzy c-means is a more suitable method to classify the FTIR spectral data in this study.

6. 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] 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.

[5] 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.

[6] 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.

7. AUTHOR PROFILE

Dr. R. Satya Prasad received Ph.D. degree in Computer Science in the faculty of Engineering in 2007 from Acharya Nagarjuna University, Andhra Pradesh. He received gold medal from Acharya Nagarjuna University for his out standing performance in Masters Degree. He is currently working as Associate Professor and H.O.D, in the Department of Computer Science & Engineering, Acharya Nagarjuna University. His current research is focused on Software Engineering. He has published 135 papers in National & International Journals. So far 20 Ph.D’s awarded under his guidance.

Marri. Suneetha working as Associate professor, Department of Computer Science, Rajah R S R K Ranga Rao College,BOBBILI, Andhra Pradesh. Her research interest in data mining.

R.Mahesh pursuing B.Tech in Dhanekula College of Engineering and Technology. Affiliated to JNTUK, Kakinada. His research interest in Data Mining, Software Engineering, CloudComputing.