

# Chapter 20 Novel Developments in Fuzzy Clustering for the Classification of Cancerous Cells using FTIR Spectroscopy

Xiao-Ying Wang<sup>1</sup>, Jonathan M. Garibaldi<sup>1</sup>, Benjamin Bird<sup>2</sup> and Mike W. George<sup>2</sup>

<sup>1</sup>*School of Computer Science and Information Technology, University of Nottingham, UK*

<sup>2</sup>*School of Chemistry, University of Nottingham, UK*

**Abstract** Cancer has become a major adversary to human health and the development and enhancement of techniques for use in its diagnosis and treatment have quickly become a focus of worldwide research. In recent years, Fourier transform infrared (FTIR) spectroscopy has been increasingly applied to the study of biomedical conditions and is becoming a powerful tool for determining the biochemical composition within a biological system. In order to analyse 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 cluster 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 has been shown that fuzzy clustering techniques such as fuzzy c-means can have clear advantages over crisp and probabilistic clustering methods, and have been widely used in medical diagnosis and pattern recognition. In this Chapter, we summarise the fuzzy clustering techniques which we have developed and successfully applied to the identification of cancer cells using FTIR spectroscopy in a selection of tissue samples which have kindly been provided by Derby General Hospital and Gloucestershire Royal Hospital, England, UK.

**Keywords:** Fourier transform infrared spectroscopy, multivariate clustering techniques, fuzzy clustering, fuzzy c-means, pattern recognition.

## 1 Introduction

As a major human health concern, cancer has become a focus for worldwide research. In Britain, more than one in three people will be diagnosed with cancer during their lifetime and one in four will die from the disease. The provision of more accurate diagnostic techniques might allow various cancers to be identified at an earlier stage and, hence, allow for earlier application of treatment. Histological evaluation of human tissue is the conventional way for clinicians to diagnose disease. It is relatively unchanged and remains the 'gold standard' since its introduction over 140 years ago. This process requires a clinician to remove and examine tissue from suspicious lesions within the patient's body and, subsequently through chemical preparation, to allow thin sections of the tissue to be cut. The addition of contrast inducing dyes allows a trained observer to identify morphological staining patterns within tissue and cells that are characteristic of the onset of disease. Nevertheless, some significant problems remain in traditional histology due to the subjective processes involved. These include missed lesions, perforation of samples and unsatisfactory levels of inter- and intra-observer discrepancy. This lack of a reliable tool for disease diagnosis has led to a considerable amount of interest investigating the application of a spectroscopic approach [1,2].

In recent years, Fourier transform infrared (FTIR) spectroscopy has been increasingly applied to the study of biomedical conditions, as it can permit the chemical characterisation of microscopic areas in a tissue sample [3-6]. There are two types of FTIR detection: FTIR mapping and FTIR imaging. In mapping, the infrared (IR) spectrum of the samples is collected a point at a time and many separate collections must be made to examine different areas of the tissue sample. However, in the latest imaging, the IR spectrum of the samples is acquired over a wide area in a single collection. Moreover, the imaging technique allows the collection of images in faster time with higher resolution. In comparison with conventional histology, FTIR spectroscopy has several advantages [7]:

- 1) It has the potential for fully automatic measurement and analysis.
- 2) It is very sensitive; very small samples are adequate.
- 3) It is potentially much quicker and cheaper for large scale screening procedures.
- 4) It has the potential to detect changes in cellular composition prior to such changes being detectable by other means.

In order to analyse the FTIR spectroscopic data from tissue samples, various techniques have been used previously: point spectroscopy, greyscale functional group mapping, digital staining [8]; and multivariate analysis methods, for instance principal component analysis [9]. Apart from these, multivariate clustering techniques have often been used to separate sets of unlabelled infrared spectral data into different clusters based

on their characteristics (this is an unsupervised process). By examining the underlying structure of a set of spectra data, clustering (also called unsupervised classification) can be performed such that spectra within the same cluster are as similar as possible, and spectra in different clusters are as dissimilar as possible. In this way, different types of cells may be separated within biological tissue. There are many clustering techniques that have been applied in FTIR spectroscopic analysis. These include hierarchical clustering [10-13], k-means (KM) clustering [10,14] and fuzzy c-means (FCM) clustering [9,10,13,14]. Studies have shown that, of the various clustering techniques, fuzzy clustering such as FCM can show clear advantages over crisp and probabilistic clustering methods [15]. In this Chapter, we review our new techniques in fuzzy clustering, including a simulated annealing based technique [7,16] to classify cancerous cells using FTIR spectroscopy. The aim of this ongoing work is to investigate whether infrared spectroscopy can be used as a diagnostic probe to identify early stages of cancer.

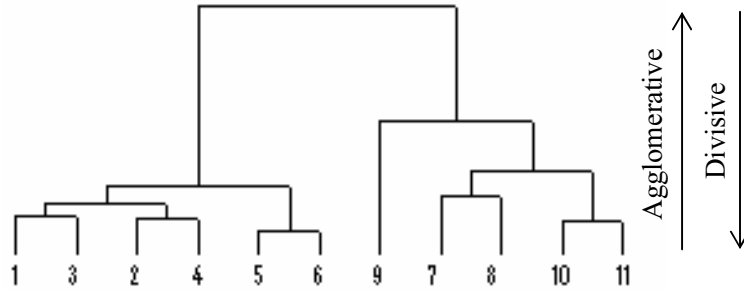
The rest of this Chapter has the following structure. In Section 2, three clustering algorithms (hierarchical, KM and FCM) often used in FTIR spectroscopic analysis are described. In Section 3, we describe the concept of cluster validity indices as a measure to evaluate the quality of clustering, and introduce the Xie-Beni cluster validity index. In the standard FCM clustering algorithm, the number of clusters has to be specified in advance. Obviously, this is not suitable if the technique is to be transferred into clinical practice. The clustering algorithm developed in Section 4 combines the simulated annealing and fuzzy c-means algorithms to achieve the automatic detection of the number of clusters (different types of tissue). Due to the complexity of the tissue sample, clustering can occasionally generate an excessive number of clusters. To counteract this problem, we propose an automatic cluster merging method in Section 5. Finally, we draw conclusions about the suitability of FTIR analysis for clustering in biological systems.

## 2 Clustering Techniques

Three unsupervised clustering algorithms that are frequently used in FTIR spectroscopy analysis are hierarchical clustering, k-means and fuzzy c-means. In this Section, these three clustering techniques are described in detail. Corresponding experiments were conducted in order to compare their performance.

### 2.1 Hierarchical Clustering

Hierarchical clustering is a way to group the data in a nested series of partitions [17]. It is a 'hard' clustering method where each datum can only belong to one cluster. Once a datum is assigned to a certain cluster, the assignment is fixed for the remainder of the process. The output of hierarchical clustering is a cluster tree, known as a *dendrogram*. It shows the similarity level between all the patterns. Figure 1 shows an example of a dendrogram in which the numbers at the bottom symbolise different data in the dataset.



**Figure 1. Hierarchical clustering dendrogram**

Based on the algorithm structure and operation, hierarchical clustering can be further categorised into agglomerative and divisive algorithms [17,18]. Agglomerative methods initially consider each datum as an individual cluster and then repeatedly merge the two closest clusters (which are measured based on the corresponding linkage method) to form a new cluster, until all the clusters are merged into one. Therefore, in an agglomerative approach, the dendrogram is generated in a bottom-up procedure. Divisive methods start by considering that all data exists in one cluster and, based on the similarity within the data, the cluster is split into two groups, such that data in the same group have the highest similarity and data in the different groups have the most dissimilarity. This process continues until each cluster only contains single datum. Hence, the divisive approach follows a top-down procedure. These two approaches are illustrated in Figure 1.

As part of an agglomerative algorithm there are many different linkage methods which provide a measure of the similarity of clusters based on the data within a cluster [19]. The main linkage methods include single link, complete link and minimum-variant algorithms (Ward's algorithm) [17,18]. Other linkages are derivatives of these three main methods. In the single link algorithm, the distance between two clusters is measured by the two *closest* data within the different clusters. By contrast, in the complete link algorithm, the distance between two clusters is measured by the two *furthest* data within the different clusters. The minimum-variant algorithm is distinct from the other two methods because it uses a variance analysis approach to measure the distance between two clusters. In general, this method attempts to minimise the sum of squares of any two hypothetical clusters which can be generated at each step. This is based on the Euclidean distance between all centroids [20]. Ward's algorithm was adopted in most of the previous applications of FTIR analysis with hierarchical clustering. In some studies (such as [10]), it produced better clustering results.

## 2.2 K-Means

K-Means (KM) clustering is a non-hierarchical clustering algorithm. It is also a ‘hard’ clustering method because the membership value of each datum to its cluster centre is either zero or one, corresponding to whether it belongs to that cluster or not. The procedure for the KM clustering algorithm can be described as follows.

Firstly let  $X = \{x_1, x_2, \dots, x_n\}$  represent a vector of real numbers, where  $n$  is the number of data.  $V = \{v_1, v_2, \dots, v_c\}$  is the corresponding set of centres, where  $c$  is the number of clusters. The aim of the K-means algorithm is to minimize the squared-error objective function  $J(V)$ :

$$J(V) = \sum_{i=1}^c \sum_{j=1}^{c_i} \|x_{ij} - v_j\|^2 \quad (1)$$

where  $\|x_{ij} - v_j\|$  is the Euclidean distance between  $x_{ij}$  and  $v_j$ .  $c_i$  is the number of data in cluster  $i$ . The  $i^{\text{th}}$  centre  $v_i$  can be calculated as follows:

$$v_i = \frac{1}{c_i} \sum_{j=1}^{c_i} x_{ij} \quad i=1 \dots c \quad (2)$$

The algorithm then proceeds as follows:

- 1) Randomly select  $c$  cluster centres.
- 2) Calculate the distance between all of the data and each centre.
- 3) Assign each datum to a cluster based on the minimum distance.
- 4) Recalculate the centre positions using Equation (2).
- 5) Recalculate the distance between each datum and each centre.
- 6) If no data were reassigned then stop, otherwise repeat from step (3).

Further details of KM clustering can be found, for example, in [21].

## 2.3 Fuzzy C-Means

FCM clustering is a fuzzy version of KM (hence, it is sometimes also called fuzzy k-means). It was proposed by Bezdek in 1981 [22]. One of differences between FCM and KM is that FCM contains an additional parameter – the ‘fuzzifier’  $m$  ( $1 < m < \infty$ ). However, as with the KM algorithm, FCM needs the number of clusters to be specified as an input parameter to the algorithm. In addition, both may suffer premature convergence to local optima [17].

The objective of the FCM algorithm is also to minimize the squared error objective function  $J(U, V)$ :

$$J(U, V) = \sum_{i=1}^n \sum_{j=1}^c (\mu_{ij})^m \|x_i - v_j\|^2 \quad (3)$$

where most of the variables are as above,  $U = (\mu_{ij})_{n \times c}$  is a fuzzy partition matrix, and  $\mu_{ij}$  is the membership degree of datum  $x_i$  to the cluster centre  $v_j$ . Parameter  $m$  is called the ‘fuzziness index’ or ‘fuzzifier’; it is used to control the fuzziness of each datum membership,  $m \in (1, \infty)$ . There is no theoretical basis for the optimal selection of  $m$ , and a value of  $m = 2.0$  is often chosen [22]. We carried out a brief investigation to examine the effect of varying  $m$  in this domain. It was found that as  $m$  increased around a value of 2.0, the clustering centres moved slightly, but the cluster assignments were not changed. However, as  $m \rightarrow \infty$ , both the fuzzy objective function  $J$  and the Xie-Beni validity index  $V_{XB}$  (see Section 3) continuously decrease ( $\rightarrow 0$ ), and the cluster assignments became unstable and further from the results of clinical analysis. Consequently, we fixed the value of  $m$  as 2.0 for all further experiments reported here.

The FCM algorithm is an iterative process to minimise Equation (3) while updating the cluster centres  $v_j$  and the memberships  $\mu_{ij}$  by:

$$v_j = \frac{\sum_{i=1}^N (\mu_{ij})^m x_i}{\sum_{i=1}^N (\mu_{ij})^m}, \forall j = 1, \dots, c \quad (4)$$

$$\mu_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{d_{ij}}{d_{ik}}\right)^{\frac{2}{m-1}}} \quad (5)$$

where  $d_{ij} = \|x_i - v_j\|$ ,  $i = 1 \dots n$  and  $j = 1 \dots c$ .

A suitable termination criterion can be that the difference between the updated and the previous  $U$  is less than a predefined minimum threshold or that the objective function is below a certain tolerance value. Furthermore, the maximum number of iteration cycles can also be a termination criterion. In our experiments we used a minimum threshold of  $10^{-7}$  and a maximum number of iterations of 100.

## 2.4 Application to FTIR Spectroscopy Clustering

Initially, we applied these three clustering techniques to FTIR spectral data sets obtained from previous clinical work provided by Chalmers et al [23] in order to compare their performance. Seven sets of FTIR data, containing tumour (abnormal), stroma (connective tissue), early keratinisation and necrotic specimens, were taken from three oral cancer patients. Parallel sections were cut and stained conventionally to help identify particular regions of interest. After acquisition of the FTIR spectra, the tissue sections were also stained and examined through cytology. Some of the material in the following Section has previously appeared in [13,14].

Figure 2 (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, clearly identifiable by their light and dark coloured stains respectively. Figure 2 (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 2 (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 are shown in Figure 3.

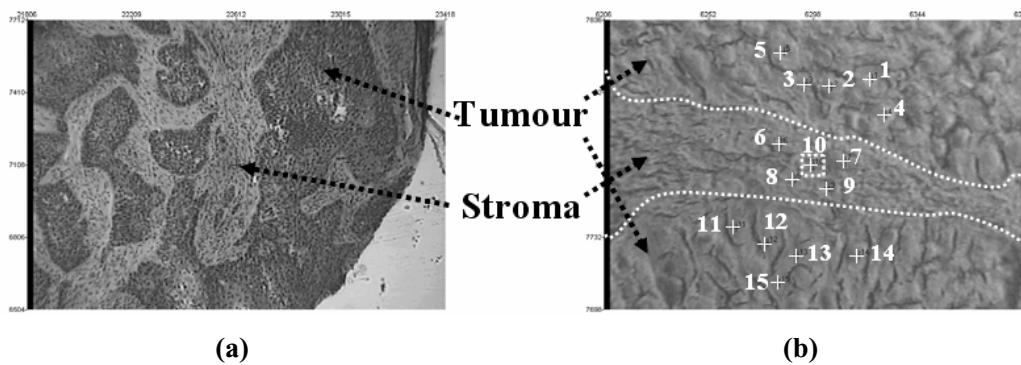


Figure 2: Tissue sample from Dataset 1; (a) 4× stained picture; (b) 32× unstained picture.

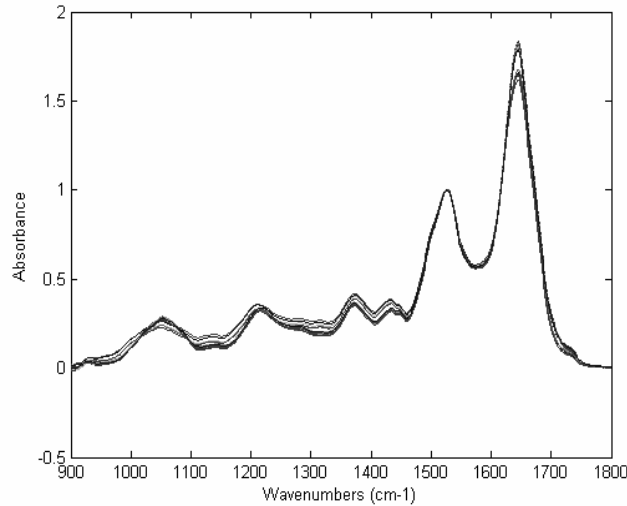


Figure 3: FTIR spectra from Dataset 1.

A Nicolet 730 FTIR spectrometer (Nicolet Instrument, Inc., Madison, USA) was used to collect the spectral data. In previous work multivariate analyses, hierarchical clustering and principal component analysis (PCA) had been applied in order to facilitate discrimination between different spectral characteristics. Firstly some *basic pre-processing* such as water vapour removal, baseline correction and normalisation had been applied. Then further pre-processing treatments were also performed empirically on the FTIR spectral data, for instance mean-centering, variance scaling and use of first-derivative spectra. Both the multivariate data analysis and pre-processing of the spectra were undertaken using Infometrix Pirouette® (Infometrix, Inc., Woodinville, WA, USA) software. The spectral range in this study was limited to a 900-1800cm<sup>-1</sup> interval. The results from multivariate analysis and cytology showed that accurate clustering could only be achieved by manually applying pre-processing techniques that varied according to the particular sample characteristics and clustering algorithms.

In our experiments, we applied the three previously mentioned clustering techniques to these seven spectral datasets obtained through conventional cytology [23], but without any extra pre-processing (i.e. only *basic pre-processing*). For hierarchical clustering, the three different types of linkage methods mentioned previously, namely ‘single’, ‘complete’ and ‘Ward’ were utilised. Due to the KM and FCM algorithms being sensitive to the initial (random) state, we ran each method 10 times. Table 1 shows the numbers of different types of tissue identified clinically (column ‘Clinical study’) and as obtained by the three clustering techniques. As mentioned previously, clustering is an unsupervised process and the results of the clustering simply 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 the clinical labelling in each case.

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

<i>Datasets names</i>	<i>Tissue types</i>	<i>Clinical study</i>	<i>Hierarchical clustering</i>			<i>KM</i>		<i>FCM</i>		
			<i>Single</i>	<i>Complete</i>	<i>Ward</i>					
Dataset 1	Tumour	10	10	10	10	10		10		
	Stroma	5	5	5	5	5		5		
Dataset 2	Tumour	10	17	9	9	9		9		
	Stroma	8	1	9	9	9		9		
Dataset 3	Tumour	8	4	8	7	3	6	4	4	
	Stroma	3	7	3	4	8	5	7	7	
Dataset 4	Tumour	12	19	12	12	11	19	13	19	11
	Stroma	7	5	7	7	8	5	6	5	8
	Early keratinisation	12	7	12	12	12	7	12	7	12
Dataset 5	Tumour	18	29	18	18	14	17	14		
	Stroma	12	1	12	12	16	13	16		
Dataset 6	Tumour	10	10	10	10	10		10		
	Stroma	5	5	5	5	5		5		
Dataset 7	Tumour	21	28	17	15	17		18		
	Stroma	14	13	18	20	18		16		
	Necrotic	7	1	7	7	7		8		

In Table 1 it can be seen that, in most of datasets, the number of data belonging to the various categories do not exactly match the results from clinical analysis. This is because some of the data that should be classified in the tumour cluster has been misclassified into the stroma cluster and vice versa. For example, in data set 2, using the hierarchical clustering single linkage method, the number of data considered as tumour is 17, while 1 is considered as stroma. We regard the extra data from these clustering techniques as the number of disagreements of classification in comparison to clinical analysis. The results of such comparison are shown in Table 2.

Table 2: Comparison results based on the number of disagreements between clinical study and clustering results.

<i>Datasets names</i>	<i>Tissue types</i>	<i>Hierarchical clustering</i>			<i>KM</i>			<i>FCM</i>		
		<i>Single</i>	<i>Complete</i>	<i>Ward</i>						
Dataset 1	Tumour	0	0	0	0	0	0	0	0	0
	Stroma	0	0	0	0	0	0	0	0	0
Dataset 2	Tumour	7	0	0	0	0	0	0	0	0
	Stroma	0	1	1	1	1	1	1	1	1
Dataset 3	Tumour	0	0	0	0	0	0	0	0	0
	Stroma	4	5	3	5	2	4	4	4	4
Dataset 4	Tumour	7	3	3	3	7	3	3	7	7
	Stroma	5	3	3	4	5	2	4	5	5
	Early keratinisation	0	0	0	0	0	0	0	0	0
Dataset 5	Tumour	12	0	0	0	0	0	0	0	0
	Stroma	1	0	0	4	1	4	4	4	4
Dataset 6	Tumour	0	0	0	0	0	0	0	0	0
	Stroma	0	0	0	0	0	0	0	0	0
Dataset 7	Tumour	7	0	0	0	0	0	0	0	0
	Stroma	0	4	6	4	4	2	2	2	2
	Necrotic	1	0	0	0	0	1	1	1	1

After running each clustering technique 10 times, it can be seen that the KM and FCM 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. It can be seen from Tables 1 and 2 that KM exhibits more variation (in 3 out of 7 datasets) than FCM (in 1 out of 7 datasets). The corresponding frequency of the differing clustering partitions obtained (out of 10 runs) is shown in Table 3.

Table 3 Clustering variations for KM and FCM within three datasets

<i>Datasets names</i>	<i>KM</i>			<i>FCM</i>	
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	-	

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

	<i>Hierarchical clustering</i>			<i>KM</i>	<i>FCM</i>
	<i>Single</i>	<i>Complete</i>	<i>Ward</i>		
<i>Average Number of Disagreements</i>	44	16	16	18.8	19.5

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, while the complete linkage and Ward methods perform best overall. However, hierarchical clustering techniques are computationally expensive (proportional to  $n^2$ , where  $n$  is the number of spectral data) and so are not suitable for very large datasets. KM and FCM have fairly good performance and, for both, the computational effort is approximately linear with  $n$ . Hence, compared with hierarchical clustering, these techniques will be far less time-consuming on large datasets [10]. Moreover, although KM has a slightly better performance than FCM (slightly fewer disagreements, on average), it can be seen from Table 3 that KM exhibits far more variation in its results than FCM. Hence, the overall conclusion was that FCM is the most suitable clustering method in this context.

### 3 Cluster Validity

Clustering validity is a concept to evaluate how good clustering results are. There are many cluster validity indices that have been proposed in the literature for evaluating fuzzy and other clustering techniques. Indices which only use the membership values such as the partition coefficient and partition entropy [24] have the advantage of being easy to compute, but are only useful for a small number of well-separated clusters. Furthermore, they also lack direct connection to the geometrical properties of the data. In order to overcome these problems, Xie and Beni defined a validity index which measures both compactness and separation of clusters [25]. In our study, we selected the Xie-Beni ( $V_{XB}$ ) cluster validity index to measure the clustering quality as it has been frequently used in recent research [26] and has also been shown to be able to detect the correct number of clusters in several experiments [27]. Some of the material in the following Section has previously appeared in [7,16].

### 3.1 Xie-Beni Validity Index

The Xie-Beni validity index,  $V_{XB}$ , can be considered as a combination of two parts. The first part is to estimate the compactness of data in the same cluster and the second part is to evaluate the separation of data in different clusters. Let  $\pi$  represent the compactness and  $s$  be the separation of the clusters. The Xie-Beni validity index can be expressed as:

$$V_{XB} = \frac{\pi}{s} \quad (6)$$

where

$$\pi = \frac{\sum_{j=1}^c \sum_{i=1}^n \mu_{ij}^2 \|x_i - v_j\|^2}{n} \quad (7)$$

and  $s = (d_{\min})^2$ .  $d_{\min}$  is the minimum distance between cluster centres, given by  $d_{\min} = \min_{ij} \|v_i - v_j\|$ . From these expressions, it can be seen that smaller values of  $\pi$  indicate the clusters are more compact, while larger values of  $s$  indicate the clusters are well separated. As a result, a smaller value of  $V_{XB}$  means that the clusters have greater separation from each other and the data are more compact within each cluster.

### 3.2 Application in FTIR Spectral Clustering Analysis

As mentioned in Section 2.3, for the FCM algorithm, the number of clusters has to be specified *a priori*, which is obviously not suitable for real-world applications. However, with the use of cluster validity indices, it is possible to discover the ‘optimal’ number of clusters within a given dataset [28]. This can be achieved by evaluating all of the possible clusters with the validity index; the optimal number of clusters can be determined by selecting the minimum value (or maximum, depending on the validity index used) of the index. The procedure of this can be described by a *FCM Based Selection Algorithm (FBSA)* [15,28,29].

This algorithm is based on the standard FCM clustering method whereby  $c_{\min}$  and  $c_{\max}$  represent the minimal and maximal number of clusters that the dataset may contain. The best data structure (C) is returned, based on the optimal cluster validity index value. It can be described in the following steps:

- 1) Set  $c_{\min}$  and  $c_{\max}$ .
- 2) For  $c = c_{\min}$  to  $c_{\max}$ 
  - 2.1) Initialise the cluster centres.
  - 2.2) Apply the standard FCM algorithm and obtain the new centre and new fuzzy partition matrix.

- 2.3) After FCM reaches its stop criteria, calculate the cluster validity (e.g.  $V_{XB}$ ).
- 3) Return the best data structure (C), which corresponds to the optimal cluster validity value (e.g. minimal  $V_{XB}$ ).

## 4 Simulated Annealing Fuzzy Clustering Algorithm

There have been many clustering methods that have been developed in an attempt to automatically determine the optimal number of clusters. Recently, Bandyopadhyay proposed a *Variable string length Fuzzy Clustering using Simulated Annealing* (VFC-SA) algorithm [30]. The proposed model was based on a simulated annealing algorithm whereby the cluster validity index measure was used 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 globally optimal solutions. The Xie-Beni index ( $V_{XB}$ ) was used as the cluster validity index to evaluate the quality of the solutions. Hence this VFC-SA algorithm can generally avoid the limitations which exist in the standard FCM algorithm. However when we implemented this proposed algorithm, it was found that sub-optimal solutions could be obtained in certain circumstances. In order to overcome this limitation, we extended the original VFC-SA algorithm to produce the Simulated Annealing Fuzzy Clustering (SAFC) algorithm. In this Section, we will describe the original VFC-SA and the extended SAFC algorithm in detail. Most of the material in this Section has appeared previously in [7,16].

### 4.1 VFC-SA Algorithm

In this algorithm, all of the cluster centres were encoded using a variable length string to which simulated annealing was applied. At a given temperature, the new state (string encoding) was accepted with a probability:  $1/(1 + \exp(-(E_n - E_c)/T))$ , where  $E_n$  and  $E_c$  represents the new energy and current energy respectively.  $T$  is the current temperature.

The  $V_{XB}$  index was used to evaluate the solution quality. The initial state of the VFC-SA was generated by randomly choosing  $c$  points from the data sets where  $c$  is a integer within the range  $[c_{\min}, c_{\max}]$ . The values  $c_{\min} = 2$  and  $c_{\max} = \sqrt{n}$  (where  $n$  is the number of data points) were used, following the suggestion proposed by Bezdek in [24]. The initial temperature  $T$  was set to a high temperature  $T_{\max}$ . A neighbour of the solution was produced by making a random alteration to the string describing the cluster centres (as described below) and then the energy of the new solution was calculated. The new solution was kept if it satisfied the simulated annealing acceptance requirement. This process was repeated for a certain number of iterations,  $k$ , at the given temperature. A cooling rate,  $r$ , where  $0 < r < 1$ , decreased the current temperature by  $T = rT$ . This was repeated until  $T$  reached the termination criteria temperature  $T_{\min}$ , at which point the

current solution was returned. The whole VFC-SA algorithm process is summarised in the following steps:

- 1) Set parameters  $T_{\max}, T_{\min}, c, k, r$ .
- 2) Initialised the string by randomly choosing  $c$  data points from the data set to be cluster centres.
- 3) Compute the corresponding membership values using Equation (5).
- 4) Calculate the initial energy  $E_c$  using  $V_{XB}$  index from Equation (6).
- 5) Set the current temperature  $T = T_{\max}$ .
- 6) while  $T \geq T_{\min}$ 
  - 6.1) For  $i = 1$  to  $k$ 
    - 6.1.1) Alter a current centre in the string.
    - 6.1.2) Compute the corresponding membership values using Equation (5).
    - 6.1.3) Compute the corresponding centres with the Equation (4).
    - 6.1.4) Calculate the new energy  $E_n$  from the new string.
    - 6.1.5) If  $E_n < E_c$  or  $E_n > E_c$  with accept probability  $>$  a random number between  $[0, 1]$ , accept the new string and set it as current string.
    - 6.1.6) Else, reject it.
  - 6.2) End for
  - 6.3)  $T = rT$ .
- 7) End while.
- 8) Return the current string as the final solution.

The process of altering a current cluster centre (step 6.1.1) 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 sizes of clusters were used to select which cluster to affect. The size,  $C_j$ , of a cluster,  $j$ , can be expressed by:

$$|C_j| = \sum_{i=1}^n \mu_{ij}, \quad \forall j = 1, \dots, c \quad (8)$$

where  $c$  is the number of clusters.

The three functions are described below.

- a) *Perturb Centre*

A random centre in the string is selected. This centre position is then modified through addition of the change rate  $cr[d] = r \cdot pr \cdot v[d]$ , where  $v$  is the current chosen centre and  $d = 1, \dots, N$ , where  $N$  is the number of dimensions.  $r$  is a random number between  $[-1, 1]$  and  $pr$  is the perturbation rate which was set through initial experimentation as 0.007 as this gave the best trade-off between the quality of the solutions produced and time taken to achieve them. *Perturb Centre* can then be expressed as  $v_{new}[d] = v_{current}[d] + cr[d]$ , where  $v_{current}[d]$  and  $v_{new}[d]$  represent the current and new centre respectively.

#### b) *Split Centre*

The centre of the biggest cluster is chosen by using Equation (8). This centre is then replaced by two new centres which are created by the following procedure. The point with the highest membership value less than 0.5 to the selected centre is identified as the reference point  $w$ . Then the distance between this reference point and the current chosen centre is calculated using:  $dist[d] = |v_{current}[d] - w[d]|$ . Finally, the two new centres are then obtained by  $v_{new}[d] = v_{current}[d] \pm dist[d]$ .

#### c) *Delete Centre*

The smallest cluster is identified and its centre deleted from the string encoding.

### 4.2 SAFC Algorithm

When we implemented the original VFC-SA algorithm on a wider set of test cases than originally used by Bandyopadhyay [30], it was found to suffer from several difficulties. In order to overcome these difficulties, four extensions to the algorithm were developed. In this Section, the focus is placed on the extensions to VFC-SA in order to describe the proposed SAFC algorithm.

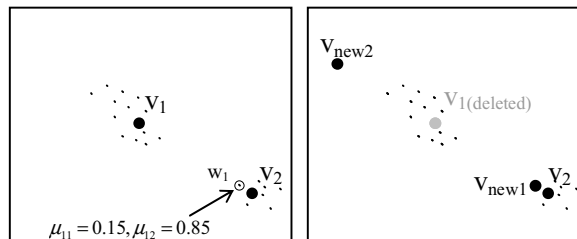
The first extension is in the initialisation of the string. Instead of the original initialisation in which random data points were chosen as initial cluster centres, the FCM clustering algorithm was applied using the random integer  $c \in [c_{min}, c_{max}]$  as the number of clusters. The cluster centres obtained from the FCM clustering are then utilised as the initial cluster centres for SAFC.

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. In contrast, if the weakest (smallest) centre is chosen, the situation in which an already good (large) centre is destabilized is avoided. Ultimately, this can lead to a quicker and more productive search as improvements to the poorer regions of a solution can be concentrated upon.

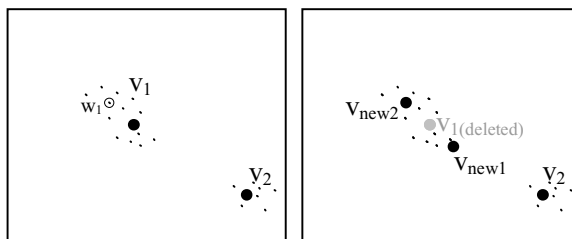
The third extension is in *Split Centre*. If the boundary between the biggest cluster and the other clusters is not obvious (not very marked), then a suitable approach is to choose a reference point with a membership degree that is less than but close to 0.5. That is to say there are some data points whose membership degree to the chosen centre is close to 0.5. There is another situation that can also occur in the process of splitting centre; the biggest cluster is separate and distinct from the other clusters. For example, let there be two clusters in a set of data points which are separated, with a clear boundary between them.  $v_1$  and  $v_2$  are the corresponding cluster centres at a specific time in the search as shown in Figure 4 (shown in two-dimensions). The biggest cluster is chosen, say  $v_1$ . Then a data point whose membership degree is closest to but less than 0.5 can only be chosen from the data points that belong to  $v_2$  (where the data points have membership degrees to  $v_1$  less than 0.5). So, for example, the data point  $w_1$  (which is closest to  $v_1$ ) is chosen as the reference data point. The new centres will then move to  $v_{new1}$  and  $v_{new2}$ . Obviously these centres are far from the ideal solution. Although the new centres are likely to be changed by the *Perturb Centre* function afterwards, it will inevitably take a longer time to ‘repair’ the solutions. In the modified approach, two new centres are created within the biggest cluster. The same dataset as in Figure 4 is used to illustrate this process. A data point is chosen,  $w_1$ , for which the membership value is closest to the mean of all the membership values above 0.5. Then two new centres  $v_{new1}$  and  $v_{new2}$  are created according the distance between  $v_1$  and  $w_1$ . This is shown in Figure 5. Obviously the new centres are better than the ones in Figure 4 and therefore better solutions are likely to be found in the same time (number of iterations).

The fourth extension is in the final step of the algorithm (return the current solution as the final solution). In the SAFC algorithm, the best centre positions (with the best  $V_{XB}$  index value) that have been encountered are stored throughout the search. At the end of the search, rather than returning the current solution, the best solution seen throughout the whole duration of the search is returned.

Aside from these four extensions, we also ensure that the number of clusters always remains within the range  $[c_{min}, c_{max}]$ . Therefore when splitting a centre, if the number of clusters has reached  $c_{max}$ , then the operation is disallowed. Similarly, deleting a centre is not allowed if the number of clusters in the current solution is  $c_{min}$ .



**Figure 4.** An illustration of *Split Centre* from the original algorithm with distinct clusters (where  $\mu_{11}$  and  $\mu_{12}$  represent the membership degree of  $w_1$  to the centres  $v_1$  and  $v_2$  respectively)



**Figure 5.** The new *Split Centre* applied to the same data set as Figure 4, above, (where  $w_1$  is now the data point that is closest to the mean value of the membership degree above 0.5)

### 4.3 Application to FTIR Spectroscopy Clustering

In order to compare the relative performance of the FCM, VFC-SA and SAFC algorithms, the following experiments were conducted. The same seven oral cancer datasets were used in this study. The number of different types of cell in each tissue section from clinical analysis was considered as the number of clusters to be referenced. This number was also used as the number of clusters for the FCM algorithm. The  $V_{XB}$  index was utilised throughout to evaluate the quality of the classification for the three algorithms. The parameters for VFC-SA and SAFC are:  $T_{\min} = 10^{-5}$ ,  $k = 40$  and  $r = 0.9$ .  $T_{\max}$  was set as 3 in all cases. That is because the maximum temperature has a direct impact on how much worse a solution can be accepted at the beginning. If the  $T_{\max}$  value is set too high, this may result in the earlier stages of the search being less productive because simulated annealing will accept almost all of the solutions and, therefore, will behave like random search. It was empirically determined that when the initial temperature was 3, the percentage of worse solutions that were accepted was around 60%. In 1996, Rayward-Smith et al discussed starting temperatures for simulated annealing search procedures and concluded that a starting temperature that results in 60% of worse solutions being accepted yields a good balance between the usefulness of the initial search and overall search time (i.e. high enough to allow some worse solutions, but low enough to avoid conducting a random walk through the search space and wasting search time) [31]. Therefore, the initial temperature was chosen based on this observation.

Solutions for the seven FTIR data sets were generated by using the FCM, VFC-SA and SAFC algorithms. Ten runs of each method were performed on each data set. As mentioned at the beginning of this Section, the number of clusters was predetermined for FCM through clinical analysis. The outputs of FCM (centres and membership degrees)

were then used to compute the corresponding  $V_{XB}$  index value. VFC-SA and SAFC automatically found the number of clusters by choosing the solution with the smallest  $V_{XB}$  index value. Table 5 shows the average  $V_{XB}$  index values obtained after 10 runs of each algorithm (best average is shown in bold).

**Table 5** Average of the  $V_{XB}$  index values obtained when using the FCM, VFC-SA and SAFC algorithms.

Dataset	Average $V_{XB}$ Index Value		
	FCM	VFC-SA	SAFC
1	0.048036	0.047837	<b>0.047729</b>
2	0.078896	0.078880	<b>0.078076</b>
3	0.291699	0.282852	<b>0.077935</b>
4	0.416011	0.046125	<b>0.046108</b>
5	0.295937	0.251705	<b>0.212153</b>
6	0.071460	0.070533	<b>0.070512</b>
7	0.140328	0.149508	<b>0.135858</b>

In Table 5, it can be seen that in all of these seven data sets, the average  $V_{XB}$  values of the solutions found by SAFC are smaller than both VFC-SA and FCM. This means that the clusters obtained by SAFC have, on average, better  $V_{XB}$  index values than the other two approaches. Put another way, it may also indicate that SAFC is able to escape sub-optimal solutions better than the other two methods.

In the data sets 1, 2, 4 and 6, the average of  $V_{XB}$  index values in SAFC is only slightly smaller than that obtained using VFC-SA. Nevertheless, when a Mann-Whitney test [32] was conducted on the results of these two algorithms, the  $V_{XB}$  index for SAFC was found to be statistically significantly lower (with  $p < 0.01$ ) than that for VFC-SA for all data sets.

The number of clusters obtained by VFC-SA and SAFC for each dataset is presented in Table 6. The numbers in parentheses indicate the number of runs for which that particular cluster number was returned. For example on dataset 5, the VFC-SA algorithm found 2 clusters in 5 runs and 3 clusters in the other 5 runs. The number of clusters identified by clinical analysis is also shown for comparative purposes.

**Table 6.** Comparison of the number of clusters obtained by clinical analysis, VFC-SA and the SAFC methods.

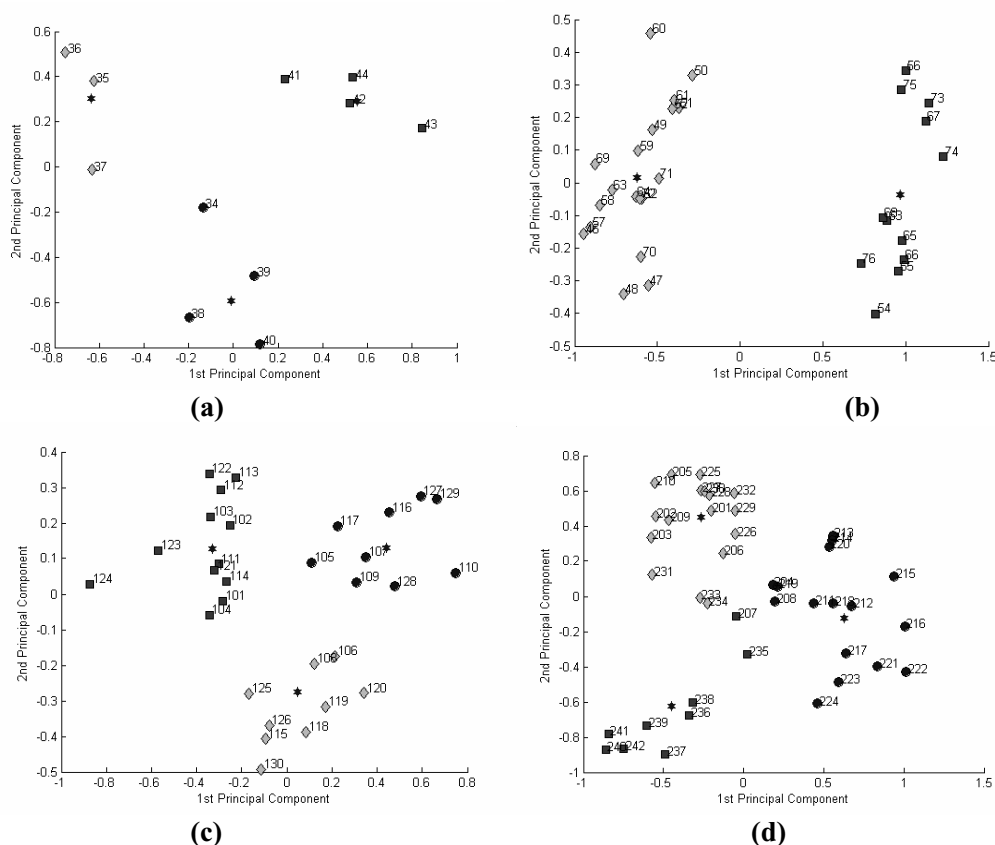
Dataset	Number of Clusters obtained		
	Clinical	VFC-SA	SAFC
1	2	2(10)	2(10)
2	2	2(10)	2(10)
3	2	2(10)	3(10)
4	3	2(10)	2(10)
5	2	2(5), 3(5)	3(10)
6	2	2(10)	2(10)
7	3	3(9), 4(1)	3(10)

In Table 6, it can be observed that in data sets 3, 4, 5 and 7, either one or both of VFC-SA and SAFC obtain solutions with a differing number of clusters than provided by clinical analysis. In fact, with data sets 5 and 7, VFC-SA produced a variable number of clusters within the 10 runs. Returning to the  $V_{XB}$  index values of Table 5, it was found that all the average  $V_{XB}$  index values obtained by SAFC are better.

It can be observed that the average  $V_{XB}$  index obtained by SAFC is much smaller than that of FCM for data sets 3 and 4. These two data sets are also the data sets in which SAFC obtained a different number of clusters to clinical analysis. In data set 3, the average  $V_{XB}$  index value for SAFC is also much smaller than for VFC-SA. This is because the number of clusters obtained from these two algorithms is different (see Table 6). Obviously a different number of clusters lead to a different cluster structure, and so there can be a big difference in the validity index. In data sets 5 and 7, the differences of  $V_{XB}$  index values are noticeable, though not as big as data sets 3 and 4.

In order to examine the results further, the data has been plotted using the first and second principal components in two dimensions. These have been extracted using the principal component analysis (PCA) technique [33,34]. The data has been plotted in this way because, although the FTIR spectra are limited to within  $900\text{cm}^{-1} - 1800\text{cm}^{-1}$ , there are still 901 absorbance values corresponding to each wave-number for each datum. The first and second principal components are the components that have the most variance in the original data. Therefore, although the data is multidimensional, the principal components can be plotted to give an approximate visualisation of the solutions that have been achieved. Figures 6 (a)-(d) show the results for data sets 3, 4, 5 and 7 respectively using SAFC (the data in each cluster is depicted using different markers and each cluster centre is presented by a star). The first and second principal components in data sets 3, 4,

5 and 7 contain 89.76, 93.57, 79.28 and 82.64 percent of the variances in the original data, respectively.



**Figure 6.** SAFC cluster results for data sets (a) 3; (b) 4; (c) 5; and (d) 7.

There are three possible explanations for the differences between the clustering results and clinical analysis. Firstly, the clinical analysis *may not* be correct – this could potentially be caused by the different types of cells in the tissue sample not being noticed by the clinical observers or the cells within each sample could have been mixed with others. Secondly, it could be that although a smaller  $V_{XB}$  index value was obtained, indicating a ‘better’ solution in technical terms, the  $V_{XB}$  index is not accurately capturing the real validity of the clusters. Put another way, although the SAFC finds the better solution in terms of  $V_{XB}$  index, this is not actually the best set of clusters in practice. A third possibility is that the FTIR spectroscopic data has not extracted the required information necessary in order to permit a correct determination of cluster numbers – i.e. there is a methodological problem with the technique itself. None of these explanations of

the difference between SAFC and VFC-SA algorithms detracts from the fact that SAFC produces better solutions in that it consistently finds better (statistically lower) values of the objective function ( $V_{XB}$  index).

## 5 Automatic Cluster Merging Method

In Section 4, it can be seen that SAFC algorithm performed well on the seven datasets. However, these are relatively small and, as the size of the dataset increases, SAFC will become time-consuming. For this reason, when large datasets are analysed, an FCM Based Selection Algorithm (FBSA – see Section 3.2) can be used to find the optimal number of clusters. In our studies, it has been found that both SAFC and FBSA occasionally identified an excessive number of clusters. This was partly due to the FCM algorithm and partly due to the cluster validity index, where all distances between data points and cluster centres are calculated using the Euclidean distance. This means that when the shapes of the clusters were significantly different from spherical, the clustering and validity measures were not effective. In addition, the complexity and range of different cell types (e.g. healthy, pre-cancerous and mature cancer) may also result in an excessive clusters being identified. Nevertheless, at this stage of study, we only focus on grouping the cells with the same clinical diagnosis into one cluster, so that the main types of the tissue can be explored through further clinical analysis. In order to achieve this, it is required to combine the most similar clusters together, such as within the suspected pre-cancerous and mature cancer cell types. Although they represent different stages of cancer, they may exhibit certain properties that make them similar to one another. This information may be contained in the existing infrared spectra. Most of material in this Section has previously appeared in [29].

A new method is proposed that will enable similar separated clusters, produced via the FBSA, to be merged together into one cluster. In order to achieve this, we again used the  $V_{XB}$  index as the measure of cluster quality.

A set of infrared spectra from one tissue section were clustered and the best data structure obtained with the corresponding minimal value of cluster validity ( $V_{XB}$ ). As mentioned above, the FCM algorithm can occasionally identify an excessive number of clusters in comparison with clinical analysis. Based on this problem, a method that can find and merge two similar clusters has been developed. In the rest of this Section, the following questions will be answered: “Given a set of  $c$  clusters, which two clusters are the most similar?” and “How should they be merged?” (specifically, how to calculate the new cluster centre).

In this study, two large lymph node cancer datasets were used. There are 821 absorbance values corresponding to each wave-number for each datum in the given infrared spectral datasets. In order to reduce the number of variables in the clustering analysis, we initially identified the first 10 principal components (PCs) that were extracted

using PCA. Within these first 10 PCs, the majority of the variance (approximately 99%) is represented from the original data. The first 2 PCs incorporate around 80% of the variance in the original data. Thus by plotting the original data in these two PC dimensions, the approximate data distribution can be visualised.

### 5.1 Finding the Two Most Similar Clusters

Previously, many algorithms have been proposed for merging clusters, for instance [20, 35]. The different approaches can generally be divided into two groups. The first group are those that select the clusters which are ‘closest’ to each other [20], and the second those that choose the ‘worst’ two clusters (judging by some cluster validity function) [35]. However, neither of these is suitable for solving the problem described in this research, illustrated in Figure 7.

A set of infrared spectral data were plotted in first and second PCs after applying the FCM based model selection algorithm. Four clusters (C1, C2, C3, and C4) were formed as shown in Figure 7. In clinical analysis, C1 and C3 are all cancer cells although they were taken from different areas of the tissue section that may contain different stages of cancer. C2 and C4 are normal nodal and reticulum cells respectively. Obviously then, the two similar clusters that need to be merged together are C1 and C3. Referring to the two types of techniques that merge clusters, the closest two clusters in the data set are C1 and C2 (see the distances between each cluster centre). Normally, a good cluster is defined by the data points within the cluster being tightly condensed around the centre (compactness). In the sample data set, clusters C1 and C2 are more compact than the other two, so the worst two clusters are C3 and C4. So, neither technique chooses the desired two clusters, C1 and C3.

In order to tackle this problem, we examined the original infrared spectra rather than searching for a relationship using the data structure in the PCA plot. Plotting the mean spectra from the separate clusters allowed the major differences between them to be more clearly visualized. The similarity between clusters was more obvious at the wavelength where the biggest difference between any two mean spectra was located. Based on this observation, the proposed method to find two similar clusters was:

- 1) Obtain the clustering results from the FCM based model selection algorithm.
- 2) Calculate the mean spectra  $\bar{A}_i$  for each cluster,

$$\bar{A}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} A_{ij} \quad (i=1\dots c) \quad (9)$$

where  $N_i$  is the number of data points in the cluster  $i$ ,  $A_{ij}$  is the absorbance of the spectrum for each data point  $j$  in cluster  $i$ , and  $c$  is the number of clusters. The size of  $\bar{A}_i$  is  $p$ , the number of wave-numbers in each spectrum (each mean spectrum is a vector of  $p$  elements).

- 3) Compute the vector of pair-wise squared differences  $D_{ij}$  between all mean spectra,

$$D_{ij} = (\overline{A}_i - \overline{A}_j)^2 \quad (i=1\dots c, j=1\dots c) \quad (10)$$

- 4) Find the largest single element,  $d_{max}$ , within the set of vectors  $D$ .  
 5) Determine the wave-number corresponding to the maximal element  $d_{max}$ .  
 6) At the wave-number identified in step (5), find the two mean spectra with minimal difference. The clusters corresponding to these spectra are merged.

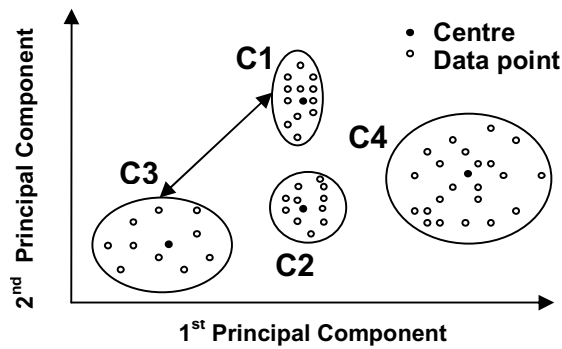


Figure 7 A set of clusters obtained from an infrared data set, plotted against the first and second principal components.

The mean spectra for the four clusters are displayed in Figure 8. We calculated the set of differences,  $D$ , between each pair of mean spectra using Equation (10). The largest difference  $d_{max}$  exists between C1 and C4 as shown in Figure 9. The wave-number that corresponds to  $d_{max}$  is  $2924 \text{ cm}^{-1}$ . Examining the four absorbance values at this wave-number we can clearly see that the two closest spectra are those belonging to clusters C1 and C3. Hence these are the two clusters selected to be merged.

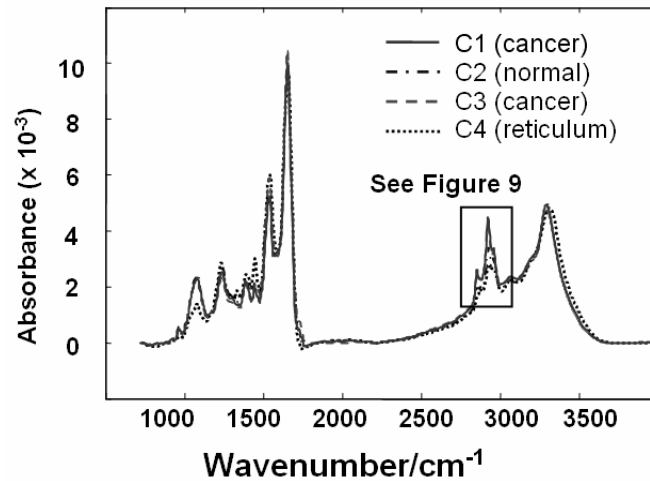


Figure 8: Example of mean infrared spectra obtained from different clusters.

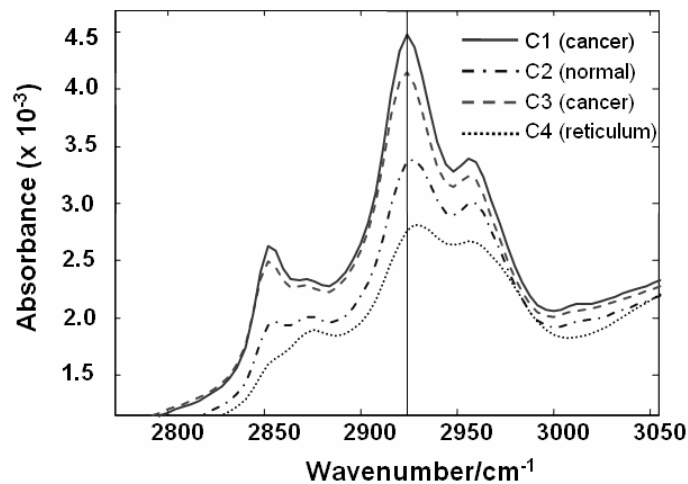


Figure 9: Enlarged region of Figure 8

## 5.2 Automatic Cluster Merging

After finding the two most similar clusters, the next stage is to merge them together. Firstly, all the data points within each of the clusters are assigned to their centre with a membership ( $\mu$ ) of one. Note that the two chosen clusters outlined above (C1 and C3) are now considered as one cluster. Data points within a particular cluster are assigned a

membership value of zero to other cluster centres. The following was used to calculate the centres [22]:

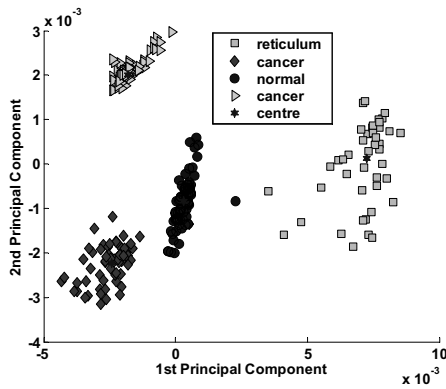
$$v_j = \frac{\sum_{i=1}^n \mu_{ij} x_i}{\sum_{i=1}^n \mu_{ij}}, \forall j = 1, \dots, c \quad (11)$$

where  $c$  is the number of clusters and  $n$  is the number of data points.

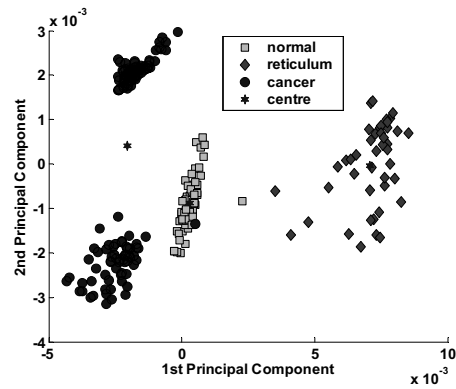
### 5.3 Application in FTIR Spectroscopic Clustering Analysis

In these experiments, we initially analysed two lymph node tissue sections named LNII5 and LNII7. The *FBSA* algorithm was firstly used to generate initial clusters with good data structure. These clusters were then further combined using the new cluster merging method. Due to the fact that different initialization states may lead to different clustering results, we ran the *FBSA* algorithm 10 times on both datasets. The results were fairly stable, and are shown in Figures 10-11.

Figures 10(a) and 11(a) show the initial clustering results obtained from the *FBSA* algorithm for LNII5 and LNII7, respectively, while Figures 10(b) and 11(b) show the clusters obtained after merging. The results clearly show that the separate cancer clusters have now been correctly merged using the new approach. In order to verify the cluster merging algorithm, the method was further applied to the oral cancer data sets described earlier, in which the SAFC clustering algorithm had obtained 3 clusters, rather than the 2 found by histological analysis. The corresponding results are shown in Figures 12 and 13.



(a)



(b)

Figure. 10: (a) LNII5 clustering results obtained from the *FBSA* algorithm. (b) LNII5 merged clusters results.

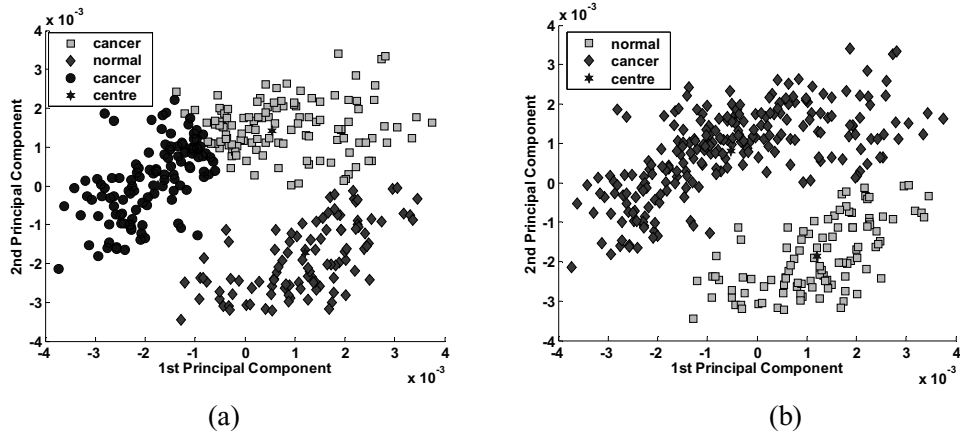


Figure. 11: (a) LNII7clutering results obtain from *FBSA* algorithm. (b) LNII7 merged clusters results.

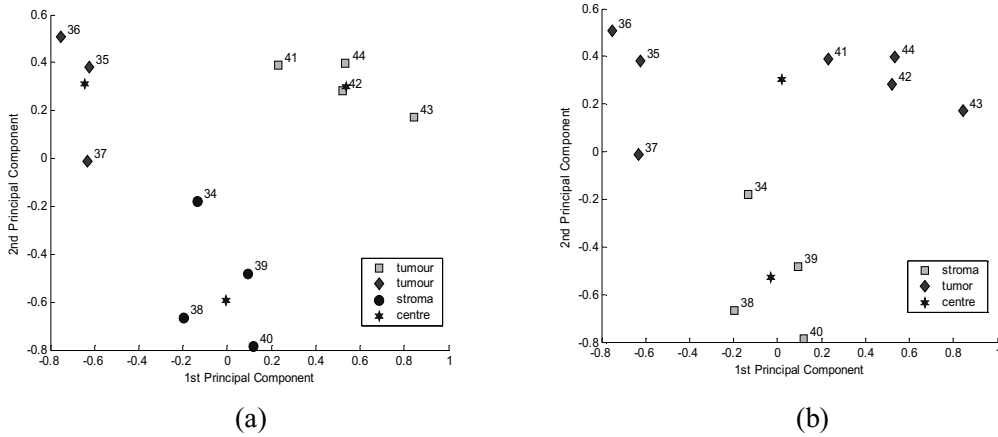


Figure. 12: (a) Dataset 3 clustering results obtain from *SAFC* algorithm. (b) Dataset 3 merged clusters results.

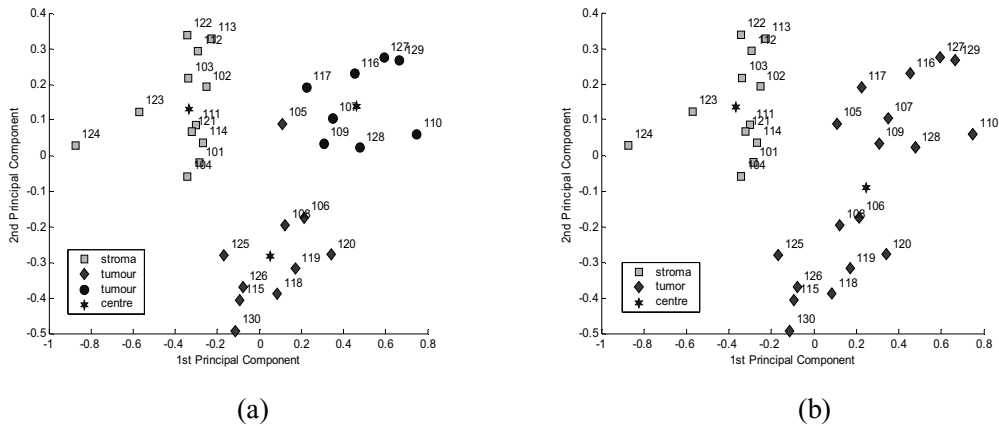


Figure. 13: (a) Dataset 5 clustering results obtain from *SAFC* algorithm. (b) Dataset 5 merged clusters results.

Altogether then, the proposed method was applied to four separate infrared spectral datasets (for which previous approaches could not obtain the correct number of clusters). For each dataset, the proposed method identified clusters that best matched clinical analysis. It should be noted that after merging clusters, there were still some misclassified data points (approximately 3-4 for LNII7 and 1-2 for the remaining datasets). However, overall the clustering accuracy was significantly improved.

## 6 Conclusion

In this Chapter, we have summarised our recent work on the analysis of non pre-processed FTIR spectra data utilising fuzzy clustering techniques. Three frequently used methods in FTIR clustering analysis, namely hierarchical clustering, k-means and fuzzy c-means clustering were described in Section 2. Each of the clustering techniques was applied to seven datasets containing FTIR spectra taken from oral cancer cells and the clustering results were compared. It can be seen from the results that the hierarchical complete linkage and Ward methods obtained the lowest number of disagreements in comparison with clinical analysis. However, in practice, these are computationally expensive when large datasets are analysed. K-means generated good results, but the variation in clustering results obtained from multiple runs implies that it may not be generally suitable in this context. In comparison to k-means and hierarchical clustering, fuzzy c-means provided reasonable performance, did not yield much variation in results and is not time-consuming in analysing large datasets. It also obtained fairly stable results throughout our experimentation.

In Section 3, one of the most frequently used cluster validity indices, the Xie-Beni index, was introduced. As was discussed, for fuzzy c-means it is important to pre-specify the number of clusters in advance. However, with the help of such a cluster validity index, it is possible to find the ‘optimal’ number of clusters in the given datasets without it being defined before the algorithm is executed. We have termed this method the “fuzzy c-means based selection algorithm”. To achieve a similar purpose, the proposed simulated annealing fuzzy clustering combined both simulated annealing and fuzzy c-means techniques in order to automatically detect the optimal number of clusters from the datasets. Due, perhaps, to the complexity of biochemical systems, both techniques can occasionally obtain an excessive number of clusters compared to clinical diagnosis. In order to attempt to solve this problem, a newly developed cluster merging method was introduced. Experiments showed that this method can find the two clusters with the most similar biochemical characteristics and then merge them together. This merging technique resulted in much improved agreement with clinical analysis. In the future, we are trying to collect a wider source of sample data for which the number of classifications is known, from a number of clinical domains, such as cervical cancer smear test screening. Establishing the techniques necessary to develop clinically useful diagnosis tools based on the automated interpretation of FTIR spectra across a range of medical domains is the ultimate goal of this research.

### **Acknowledgements**

The authors would like to thank Professor H. Barr, Dr. N. Stone and Dr. J. Smith from the Gloucestershire Royal Hospital for their aid in sample collection and diagnosis. We are particularly grateful to Professor M. A. Chester and Mr. J. M. Chalmers for many helpful discussions. The authors would also like to thank the anonymous referees for their comments which have helped improve this paper.

### **Reference List**

- [1] Johnson KS et al 2004, Elastic scattering spectroscopy for intraoperative determination of sentinel lymph node status in the breast, *Journal of Biomedical Optics*, **9** (6), 1122-1128.
- [2] Godavarty A et al 2004, Diagnostic imaging of breast cancer using fluorescence-enhanced optical tomography: phantom studies, *Journal of Biomedical Optics*, **9** (3), 486-496.
- [3] Fernandez DC et al 2005, Infrared spectroscopic imaging for histopathologic recognition, *Nature Biotechnology*, **23** 469-474.

- [4] Lasch P et al 2002, Characterization of Colorectal Adenocarcinoma Sections by Spatially Resolved FT-IR Microspectroscopy, *Applied Spectroscopy*, **56** (1), 1-9.
- [5] Chiriboga L et al 1998, Infrared Spectroscopy of Human Tissue: IV. Detection of Dysplastic and Neoplastic Changes of Human Cervical Tissue via Infrared Microscopy, *Cellular & Molecular Biology*, **44** (1), 219-230.
- [6] Choo LP et al 1996, In situ characterization of beta-amyloid in Alzheimer's diseased tissue by synchrotron Fourier transform infrared microspectroscopy, *Biophysical Journal*, **71** (4), 1672-1679.
- [7] Wang XY and Garibaldi JM 2005, Simulated Annealing Fuzzy Clustering in Cancer Diagnosis, *Informatica*, **29** (1), 61-70.
- [8] McIntosh L et al 1999, Analysis and Interpretation of Infrared Microscopic Maps: Visualization and Classification of Skin Components by Digital Staining and Multivariate Analysis, *Biospectroscopy*, **5** 265-275.
- [9] Richter T et al 2002, Identification of Tumor Tissue by FTIR Spectroscopy in Combination with Positron Emission Tomography, *Vibrational Spectroscopy*, **28** 103-110.
- [10] Lasch P et al 2004, Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis, *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, **1688** (2), 176-186.
- [11] Romeo MJ and Diem M 2005, Infrared spectral imaging of lymph nodes: Strategies for analysis and artifact reduction, *Vibrational Spectroscopy*, **38** 115-119.
- [12] Wood BR et al 2004, Fourier transform infrared (FTIR) spectral mapping of the cervical transformation zone, and dysplastic squamous epithelium, *Gynecologic Oncology*, **93** (1), 59-68.
- [13] Wang XY, Garibaldi JM, and Ozen T 2003, Application of The Fuzzy C-Means Clustering Method on the Analysis of non Pre-processed FTIR Data for Cancer Diagnosis, in *Proceedings of the 8th Australian and New Zealand Conference on Intelligent Information Systems*, 233-238.
- [14] Wang XY and Garibaldi JM 2005, A Comparison of Fuzzy and Non-Fuzzy Clustering Techniques in Cancer Diagnosis, in *Proceedings of second international conference in Computational Intelligence in Medicine and Healthcare The Biopattern Conference*, 250-256.

- [15] Sun H, Wang S, and Jiang Q 2004, FCM-Based Model Selection Algorithms for Determining the Number of Clusters, *Pattern Recognition*, **37** 2027-2037.
- [16] Wang XY, Whitwell G, and Garibaldi JM 2004, The Application of a Simulated Annealing Fuzzy Clustering Algorithm for Cancer Diagnosis, in *Proceedings of the IEEE 4th International Conference on Intelligent System Design and Application*, 467-472.
- [17] Jain AK, Murty MN, and Flynn PJ 1999, Data Clustering: A Review, *ACM Computing Surveys*, **31** (3), 264-323.
- [18] Jiang D, Tang C, and Zhang A 2004, Cluster Analysis for Gene Express Data: A Survey, *IEEE Transactions on Knowledge and data Engineering*, **16** (11), 1370-1386.
- [19] Garrett-Mayer E and Parmigiani G 2004, Clustering and Classification Methods for Gene Expression Data Analysis, *Johns Hopkins University, Dept. of Biostatistics Working Papers, Johns Hopkins University*, The Berkeley Electronic Press.
- [20] Ward JH 1963, Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, **58** (301), 236-244.
- [21] MacQueen JB 1967, Some methods of classification and analysis of multivariate observations, in *Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281-297.
- [22] Bezdek J 1981, *Pattern Recognition With Fuzzy Objective Function Algorithms*, New York, Plenum.
- [23] Allibone R et al , 2002, FT-IR microscopy of oral and cervical tissue samples, *Internal Report*, Derby City General Hospital
- [24] Bezdek J 1998, *Pattern Recognition in Handbook of Fuzzy Computation*, Boston, NY, IOP Publishing Ltd.
- [25] Xie XL and Beni G 1991, A Validity Measure for Fuzzy Clustering, *IEEE Trans. Pattern Analysis and Machine Intelligence.*, **13** (8), 841-847.
- [26] Kim DW, Lee KH, and Lee D 2004, On Cluster Validity Index for Estimation of the Optimal Number of Fuzzy Clusters, *Pattern Recognition*, **37** 2009-2025.

- [27] Pal NR and Bezdek J 1995, On Cluster Validity for the Fuzzy C-Means Model, *IEEE Trans.Fuzzy System.*, **3** 370-379.
- [28] Rezaee MR, Lelieveldt BPF, and ReiBer JHC 1998, A New Cluster Validity Index for the Fuzzy C-Means, *Pattern Recognition Letters*, **19** 237-246.
- [29] Wang XY et al 2005, Fuzzy Clustering in Biochemical Analysis of Cancer Cells, in *Proceedings of Fourth Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2005)*, 1118-1123.
- [30] Bandyopadhyay S 2003, Simulated Annealing for Fuzzy Clustering: Variable Representation, Evolution of the Number of Clusters and remote Sensing Applications, unpublished, private communication.
- [31] Rayward-Smith VJ and others 1996, *Modern Heuristic Search Methods*, John Wiley & Sons.
- [32] Conover WJ 1999, *Practical Nonparametric Statistics*, John Wiley & Sons.
- [33] Causton DR 1987, *A Biologist's Advanced mathematics*, London, Allen & Unwin.
- [34] Jolliffe IT 1986, *Principal Component Analysis*, New York, Springer-Verlag.
- [35] Xie Y, Raghavan VV, and Zhao X 2002, 3M Algorithm: Finding an Optimal Fuzzy Cluster Scheme for Proximity Data, **1**, 627-632.