

A novel fuzzy clustering algorithm for the analysis of axillary lymph node tissue sections

Xiao-Ying Wang · Jonathan M. Garibaldi · Benjamin Bird · Michael W. George

Published online: 16 June 2007
© Springer Science+Business Media, LLC 2007

Abstract Recently Fourier Transform Infrared (FTIR) spectroscopic imaging has been used as a tool to detect the changes in cellular composition that may reflect the onset of a disease. This approach has been investigated as a mean of monitoring the change of the biochemical composition of cells and providing a diagnostic tool for various human cancers and other diseases. The discrimination between different types of tissue based upon spectroscopic data is often achieved using various multivariate clustering techniques. However, the number of clusters is a common unknown feature for the clustering methods, such as hierarchical cluster analysis, k-means and fuzzy c-means. In this study, we apply a FCM based clustering algorithm to obtain the best number of clusters as given by the minimum validity index value. This often results in an excessive number of clusters being created due to the complexity of this biochemical system. A novel method to automatically merge clusters was developed to try to address this problem. Three lymph node tissue sections were examined to evaluate our new method. These results showed that this approach can merge the clusters

which have similar biochemistry. Consequently, the overall algorithm automatically identifies clusters that accurately match the main tissue types that are independently determined by the clinician.

Keywords FTIR imaging · Fuzzy c-means · Clustering · Automated merge clusters · Lymph node

1 Introduction

Histological evaluation of human tissue has been relatively unchanged since its initial introduction over 140 years ago. This approach remains the gold standard used by clinicians for disease diagnosis. This process involves analyzing a biopsy of suspicious lesions within the body that is chemically prepared to allow thin sections of the tissue to be cut. The addition of contrast inducing dyes allows a trained observer to identify changes within tissue and cells that are characteristic of the onset of disease. However, traditional histology remains a subjective technique and significant problems are often encountered including missed lesions and an unsatisfactory discrepancy between pathologists. These problems in diagnosis have led to considerable interest in investigating the application of spectroscopic approaches for disease diagnosis [1, 2]. Fourier transform infrared (FTIR) spectroscopic imaging provides a chemical fingerprint of microscopic areas within a tissue sample [3–6]. Contrast between different spatial areas of a sample is due to the chemical differences that are occurring within the cells of the tissue. Thus, a chemical image of the examined tissue section can be constructed, similar to that of conventional histological staining, allowing the identification of tissue pathology. Therefore an FTIR imaging approach could have several advantages over conventional histology

X.-Y. Wang (✉) · J.M. Garibaldi
School of Computer Science and Information Technology, The
University of Nottingham, Nottingham, UK
e-mail: xyw@cs.nott.ac.uk

J.M. Garibaldi
e-mail: jmg@cs.nott.ac.uk

B. Bird · M.W. George
School of Chemistry, The University of Nottingham, Nottingham,
UK

B. Bird
e-mail: pcxbb1@nottingham.ac.uk

M.W. George
e-mail: mike.george@nottingham.ac.uk

including diagnostic decisions being based upon the chemical characteristics of the tissue without the subjective analysis of a pathologist. Furthermore, the whole process could be fully automated.

The complexities of biological systems have meant that multivariate clustering techniques have often been used to analyze complex FTIR spectra from tissue sections. The clustering process is performed in order to group a set of unlabeled infrared spectra into a certain number of clusters, such that the spectra with the most similar characteristics are in the same group and the spectra with the most dissimilar characteristics are in different groups resulting in different types of tissue being separated. The multivariate clustering techniques which have been used in related areas are Hierarchical Cluster Analysis [7–9], k-means clustering [9], and fuzzy c-means clustering [9]. Of particular relevance to this work is the use of HCA to lymph node tissue sections [8]. Whilst all of these techniques have been shown to be able to group different types of tissue, the number of clusters is an unknown feature and further analysis requires human input. Real world applications would benefit from a clustering method that could automatically separate different and a varying number of tissue types from a tissue section.

The motivation behind this work is to create an automated method to analyze FTIR spectra and to separate them into a clinically meaningful number of clusters. In our previous work, a fuzzy clustering algorithm featuring simulated annealing (SAFC) was used to automatically detect the ‘optimal’ number of clusters found within a FTIR dataset. The dataset used in this study comprised of spectra that had been collected from a variety of different tissue types [10, 11]. This algorithm begins by generating a random number of clusters and then traverses the search space using three different neighborhood operations: (i) *perturb centre*, (ii) *delete centre* and (iii) *split centre*. The configuration with the minimum cluster validity index value is returned. The results show that this algorithm was able to obtain the same number of clusters as clinical analysis in four out of seven datasets. However, smaller Xie–Beni validity index values were achieved in some datasets even though the numbers of clusters were different from the clinical analysis. Although the SAFC algorithm performed well on the small datasets, it proved to be very time consuming on larger datasets. With the aim of overcoming this problem, in our latest work [12], a refined FCM based clustering algorithm was developed to find the ‘optimal’ number of clusters.

Both the SAFC and FCM based clustering algorithms can automatically detect the number of clusters based on the clustering structure and this results in the minimum validity index value. However, both algorithms occasionally identified an excessive number of clusters compared to clinical analysis. This was partly due to the FCM algorithm and cluster validity index, where all distances between data points

and cluster centres are calculated using their Euclidean distances. This means that when the shapes of the clusters were significantly different from spherical, the clustering and validity measures were not effective. However, the complexity and range of the different cell types (e.g. healthy, pre-cancerous and mature cancer) may also lead to an excessive number of clusters being identified. At this stage of our 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, we need to combine the clusters with most similar characteristics together, e.g. within the suspected pre-cancerous and mature cancer cell types, as they may exhibit similar properties to one another even though they are different stages of cancer. This information may be contained in the existing infrared spectra.

In this paper, a new method is proposed to automatically merge clusters, which identifies the two most similar clusters generated by the initial FCM based clustering algorithm and merges them. The merged cluster will then be considered as a new cluster to rejoin the rest of merging iteration until a stop criterion is reached. In these experiments, the Sun–Wang–Jiang index is used as the cluster validity index (V_{SWJ}) because it has recently been shown to obtain the best results in comparison with other existing validity indices [13]. In the following sections, firstly the feature selection and reduction methods are described followed by some background on the well-known FCM clustering algorithm, and then cluster validity indices are briefly introduced. Subsequently the FCM based clustering algorithm is presented in brief and then our proposed cluster merging method is described in detail. We evaluate the use of our new algorithm to analyze the FTIR images of three selected tissue sections. In Sect. 5, we draw conclusions about the proposed new clustering method.

2 Feature reduction

In the given infrared spectral datasets (see Sect. 7.4), there are between 5000–8000 spectra in each dataset. For each spectrum, there are 821 absorbance values, one corresponding to a data point every 4 cm^{-1} . Thus the total size of the datasets ranges from 5000×821 (4.1×10^6) to 8000×821 (6.6×10^6). The application of clustering analysis to such large datasets can be very time consuming. In addition, it is difficult to visualize the distribution of such data. In this study, principal component analysis (PCA) [14] was used to reduce the number of variables and also used to permit visualization. PCA has been widely used for the analysis infrared spectra [17–20]. PCA is a typical multivariate statistical technique that has been widely applied in the field of data analysis and compression. This technique linearly

transforms a number of correlated variables into a new set of uncorrelated variables, called principal components (PCs). PCA achieves this by rotating the original axes to produce orthogonal axes that are uncorrelated to each other [14]. The rotation procedure is a linear transformation of the original dataset and, therefore, if all the variables are included in the rotation, then all the original information is retained. Within the transformed variables, the first principal component (PC1) identifies the dimension with the maximum variation in the original data and the second (PC2) is the dimension with the second largest variation, etc. As the number of PC increases, the less variance each component contains. Therefore, the first PCs usually contain the most influential variations from the original data and it is this property that yields the major advantage of the approach by allowing the dimensionality of the data to be reduced whilst the most significant features of the data are usually retained. It is possible that the dimensions carrying most variation do not correspond to those with the most discriminate power, but this is not common.

In our experiments, the first 10 PCs of the FTIR spectra were extracted as the input values of the FCM based clustering algorithm. For the three datasets used in our investigation, the first two PCs represent, respectively, 70.9%, 89.1% and 94.1% of the variance. Increasing this to the first 10 PCs improves the representation to 95.6%, 99.1% and 99.8% of the variance in the three datasets. In order to visualize the data distribution, we can plot the original data using the first two PCs dimensions. Details on feature reduction approaches have been described previously by many authors; see, for example [21].

3 FCM clustering algorithm

The FCM algorithm, proposed by Bezdek in 1981 [22], is one of most frequently used methods in pattern recognition. It is based on minimization of the objective function (1) to achieve good classifications.

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

Consider clustering the dataset $X = \{x_1, x_2, \dots, x_n\}$ where n is the number of data points. $V = \{v_1, v_2, \dots, v_c\}$ is the set of corresponding cluster centres in the dataset X , where c is the number of clusters. μ_{ij} is the membership degree of data x_i to the cluster centre v_j . Meanwhile, μ_{ij} has to satisfy the following conditions:

$$\mu_{ij} \in [0, 1] \quad \forall i = 1, \dots, n, \quad \forall j = 1, \dots, c, \tag{2}$$

$$\sum_{j=1}^c \mu_{ij} = 1 \quad \forall i = 1, \dots, n. \tag{3}$$

$U = (\mu_{ij})_{n \times c}$ is a fuzzy partition matrix. $\|x_i - v_j\|$ denotes the Euclidean distance between x_i and v_j . Parameter m is called the ‘‘fuzziness index’’, it is used to control the fuzziness of membership of each datum. The value of m should be within the range $m \in [1, \infty]$. There is no theoretical basis for the optimal selection of m , but a value of $m = 2.0$ is usually chosen. The FCM algorithm can be performed by following steps.

- (1) Fix the number of clusters, c , where $2 \leq c < n$, and initialize fuzzy partition matrix U with a random value such that it satisfies conditions (2) and (3).
- (2) Calculate the fuzzy centres v_j using

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

- (3) Update the fuzzy partition matrix U with

$$\mu_{ij} = 1 / \sum_{k=1}^c (d_{ij} / d_{ik})^{2/m-1} \tag{5}$$

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

- (4) Repeat step (2) to (3) until one of the termination criteria is satisfied.

In step (4), termination criteria can be that the difference between updated and previous U is less than a predefined minimum threshold or that the objective function is below a certain tolerance value. Additionally, the maximum number of iteration cycles can also be a termination criterion.

4 Cluster validity

In cluster analysis, one of the most important issues is to evaluate the quality of the clustering results. Cluster validity indices are measures that somehow express the quality of the clusters, usually comprising a combination of measures taking into account the *compactness* of the clusters in comparison with the *separation* of the clusters. Many cluster validity indices have been proposed, such as the partition coefficient and partition entropy [23], Dunn’s index [24], Xie–Beni index [16], etc. In this experimentation, the Sun–Wang–Jiang index was used as the cluster validity index (V_{SWJ}) for the reason already mentioned above. The minimum of V_{SWJ} refers to the optimal value of this validity index.

A validity index can be utilized during the clustering process in order to automatically generate the number of clusters. This can be implemented by applying various clustering algorithms with a range of cluster numbers, and returning the partition with the best cluster validity index value [25]. In addition, there are many other clustering techniques which also can determine the number of clusters, such as [26, 27]. A survey of multivariate clustering techniques can be found in past literature, for instance [28, 29].

5 FCM based clustering algorithm

This FCM based clustering algorithm is a good example of using cluster validity to find out the optimal number of clusters from a given dataset. It is composed of two parts, where the first part is based on running the FCM clustering method within a certain range of number of clusters. The best data structure (C) is obtained by choosing the corresponding optimal value of cluster validity (V_{SWJ}) from all the possible clustering structures [13]. The second part of this algorithm is based on taking the best data structure from the first part, and then slightly perturbing each cluster centre for a number of iterations while the validity index for each new data structure is calculated. The data structure corresponding to the optimal validity index is then returned. In the following description of the algorithm, the minimal and maximal number of clusters is referred to as c_{min} and c_{max} .

- (1) Set c_{min} and c_{max} (in the experiments, we set $c_{min} = 2$ and $c_{max} = 10$).
- (2) For $c = c_{min}$ to c_{max}
 - (2.1) Initialize the cluster centres.
 - (2.1) Apply the standard FCM algorithm and obtain the new centre and new fuzzy partition matrix.
 - (2.3) Calculate the cluster validity (V_{SWJ}).
- (3) Obtain the good data structure (C) that corresponds to the minimal V_{SWJ} .
- (4) Set current C as the best data structure (C_{best}).
- (5) For $i = 1$ to 100
 - (5.1) Random slightly perturb the current C .
 - (5.2) Calculate the new membership value and validity index value (V_{SWJ}) corresponding to the new data structure (C_{new}).
 - (5.3) If the new V_{SWJ} is smaller than current C_{best} V_{SWJ} value, then set the C_{new} as the C_{best} , otherwise, go back to step 5.1.
 - (5.4) Set the C_{best} as current C , and go back to step 5.1.
- (6) Return the best data structure C_{best} with the minimal V_{SWJ} value.

During the clustering process, each data point is assigned to the cluster for which the degree of the fuzzy partition matrix is maximal.

6 The basis of a new automated method to merge clusters

The motivation to create a new merge clustering method was based on previous attempts at clustering spectral data. In previous work, a dataset that comprised of FTIR spectra collected from a variety of different tissue types was analyzed via SAFC and FCM based clustering [10–12]. This dataset comprised 276 individual spectra that were collected from

regions of a tissue section diagnosed as being ‘cancer’ (cancerous cortex tissue: 159 spectra), ‘normal’ (benign cortex tissue: 72 spectra) and ‘reticulum’ (fibrocollagenous tissue: 45 spectra). When the number of clusters was set at three to match the clinical analysis, the clustering results did not match the clinical diagnosis (possibly due to the Euclidean distance measurement in FCM). Some FTIR spectra from cancerous regions were incorrectly grouped with spectra taken from regions non-cancerous tissue.

When we subsequently applied the FCM based clustering algorithm, four clusters ($C1$, $C2$, $C3$ and $C4$) were obtained, as shown in Fig. 1. However, two of the four clusters corresponded to one type of tissue (cancerous). In the remaining two clusters, the majority of the data was classified into the right group (although there were two spectra data that were misclassified). The excessive number of clusters may have been caused by the fact that the FTIR spectra taken from cancerous regions were taken from diverse areas of tissue, which might have contained cells at different stages of the cancer (e.g. pre-cancerous and mature cancer). As mentioned above, at this stage, we only wish to cluster cells which have the same clinical diagnosis.

In literature some split-and-merge techniques have been used to determine the number of clusters. Iterative Self-Organizing Data Analysis (ISODATA) is a typical algorithm that uses a split-and-merge technique [30]. It starts with a large number of clusters and then, based on various split and merge cluster criteria, determines the final number of clusters. Some of criteria are, for example if the number of data points within a cluster is less than a certain threshold, or if the distance between two clusters is less than a certain threshold. Many parameters feature within the criteria,

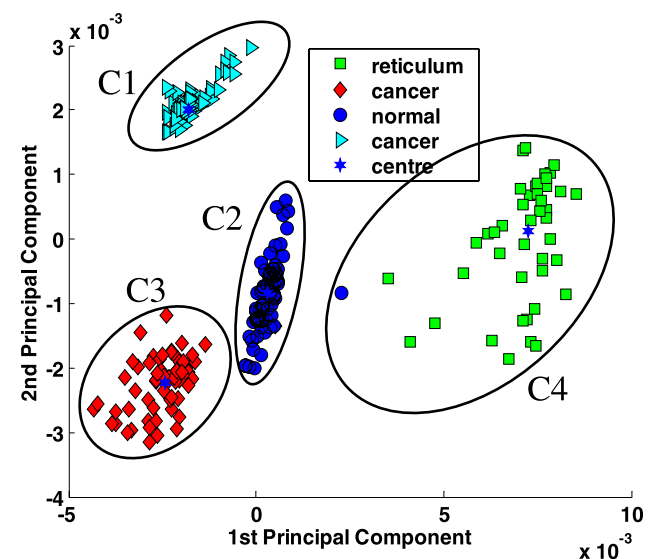


Fig. 1 An extracted spectral dataset after applying FCM based clustering algorithm

which have to be declared in advance and which may directly affect the clustering results [31].

An alternative split-and-merge clustering technique developed by Chaudhuri et al. [32] splits the clusters by observing the density in different directions and examining the number of data within different regions made up of strips. This method starts by including all the data as one cluster and then iteratively performs a split and merge process based on the number of data in the strip area which is determined by certain threshold values. Similarly to the ISODATA technique, many input parameters need to be specified before the approach can be executed. Other split-and-merge clustering algorithms can be found in [33–35]. In general, all of these techniques process the split-and-merge procedure based on the dataset itself. Other approaches to automatically determine the number of clusters have also been proposed in image processing contexts, for example [36, 37].

Several algorithms have been proposed to merge clusters (e.g. [15, 16]). Two of the most common types of methodology will be discussed by use of an example. The first method identifies and merges clusters that lie ‘closest’ to each other [15] in multi-dimensional space. In contrast, the second method identifies the ‘worst’ clusters and merges them together. This is achieved by use of a cluster validity function, the most common of which measure the compactness of the defined clusters [16]. Informally, a ‘good’ cluster is defined by the property that data points within the cluster are tightly condensed around the centre (high compactness). When applying these methods to the dataset shown in Fig. 1, the two closest clusters are C2 and C3 (see the distance between each cluster centre). In this dataset, C1 and C2 are more compact than the other two and the worst two clusters are C3 and C4. However, the two clusters that should be merged together are C1 and C3 (both are collected from cancer tissue). Hence, neither of these approaches for merging clusters is suitable for solving the problem here.

An alternative solution was developed based on examining the original infrared spectra rather than searching for a relationship using the clustering structures in the PCA space. Plotting the mean spectra from the separate clusters allows the major differences between them to be more clearly visualized. The similarity between clusters is more obvious at the wavenumber corresponding to the IR frequency that provides the largest variance between spectra. Our proposed automated merge clustering method is based on this observation and can be divided into two main stages. The first stage is to identify the frequency at which the greatest variance between mean spectra is observed. The second step is to repeatedly determine the most similar clusters and merge them, until certain termination criterion has been reached. In the following section we describe these two steps in detail.

Step 1: Determine a reference frequency

The reference frequency is defined as the frequency at which the biggest difference between any two mean spectra is found. The full procedure of determining this frequency is:

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

$$\overline{A}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} A_{ij} \quad (i = 1, \dots, c) \quad (6)$$

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 ; c is the number of clusters. The size of \overline{A}_i is p , the number of data points in each spectrum (each mean spectrum is a vector of p elements).

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

$$D_{ij} = |\overline{A}_i - \overline{A}_j| \quad (i = 1, \dots, c, j = 1, \dots, c). \quad (7)$$

- (4) Find the largest single element, d_{\max} , within the set of vectors D .
- (5) Determine the wavenumber corresponding to the maximal element d_{\max} .

The mean spectra obtained for four clusters are displayed in Fig. 2. We calculated the set of differences, D , between each pair of mean spectra using (7). The largest difference d_{\max} exists between C1 and C4, as shown in Fig. 3. The IR frequency that corresponds to d is 2924 cm^{-1} .

Step 2: Automatically merging clusters

The next step is to find the most similar clusters and then to merge them together. This is determined of the use of the absorbance intensity for each mean spectrum at the reference

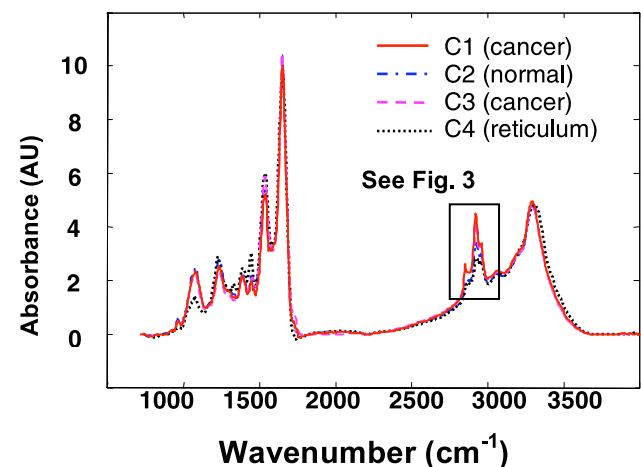


Fig. 2 Mean infrared spectra obtained from different clusters

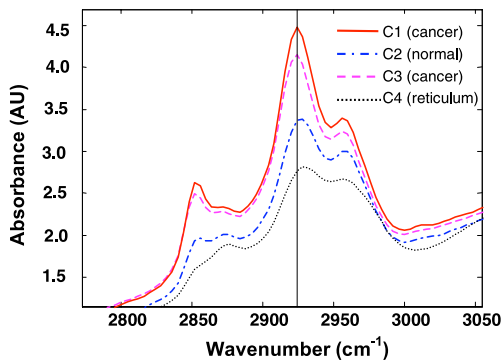


Fig. 3 Enlarged region of Fig. 2

frequency. Clusters are therefore merged dependent upon similarities in their IR spectra rather than clustering structure in multivariate space. As this is an iterative process, the merging procedure will end when at least one of the termination criteria has been satisfied. Assume currently there are C mean spectra. The detailed information can be described as below:

- (1) Obtain C absorbance values of mean spectra at reference frequency from step 1, re-sort them in ascending order.
- (2) Calculate the distance $dist$ between these sorted and adjacent absorbance values (note that the size of $dist$ now is $C - 1$).
- (3) Pick up the smallest distance $dist_{min}$ and find out the two most similar clusters which correspond to this distance.
- (4) Merge these two clusters if they satisfy the merging condition: $dist_{min} \leq \text{average of rest of } dist$ (without $dist_{min}$). The average absorbance value for the two merge clusters is then calculated and is considered as a new object to join the rest of merging iteration. Go back to (1).
- (5) When there are only two $dist$ remaining, merge the two clusters which correspond the $dist_{min}$ if the following merging condition is satisfied: $dist_{min} \leq 1/2 \text{ rest of } dist$ OR $(dist_{min} - 1/2 \text{ rest of } dist)/dist_{min} \leq 0.1$. Again, the average of these two mean spectra absorbances is considered as a new object to replace them.
- (6) The merging process stops if there are only two clusters left or no merging conditions are satisfied.

In this section, we use the same example used above to illustrate this procedure. In Fig. 4, \bar{A}_i ($i = 1, \dots, 4$) is the mean absorbance values from the normalized spectra of each obtained cluster are $\bar{A}_1 = 0.0045$, $\bar{A}_2 = 0.0034$, $\bar{A}_3 = 0.0041$ and $\bar{A}_4 = 0.0028$ respectively. The line corresponds to the reference frequency 2924 cm^{-1} . After sorting \bar{A}_i in ascending order, their new arrangement is $\bar{A}_4, \bar{A}_2, \bar{A}_3$ and \bar{A}_1 . The distance between the average absorbance intensities is described as follows $Dist = \{d_1, d_2, d_3\}$. It is then trivial to calculate $d_1 = 0.0006$, $d_2 = 0.0007$, $d_3 = 0.0004$, which is the $dist_{min}$. The average of rest of $dist = (0.0006 +$

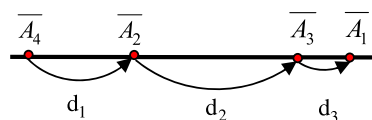


Fig. 4 Four mean spectra absorbance values at the reference frequency

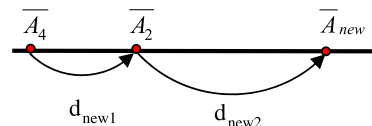


Fig. 5 After merging two most similar clusters, the resorted absorbance distribution

Fig. 6 The merging situation when there are two dist left (type 1)

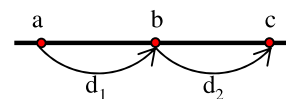
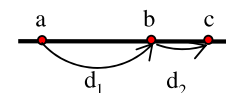


Fig. 7 The merging situation when there are two dist left (type 2)



$0.0007)/2 = 0.00065$ is greater than $dist_{min}$, this satisfies the merging condition in (4). Therefore, the two clusters which correspond to $dist_{min}$ (C_1 and C_3) are merged together. After this, the average of these two clusters' mean spectra absorbance intensities $\bar{A}_{new} = 0.0043$ replaces these values. Re-sort the new array of the mean spectra absorbance intensities to, \bar{A}_4, \bar{A}_2 and \bar{A}_{new} , as displayed in Fig. 5. The corresponding new distances are $d_{new1} = 0.0006$ and $d_{new2} = 0.0009$. Reference (5), $dist_{min}$ (0.0006) is not smaller than or equal to $1/2 \text{ rest of } dist$ (0.00045); meantime, it does not satisfy the second condition either. Hence, in this situation, no merging conditions are satisfied, and so the iteration stops as defined in (6).

As may be noticed, the merging condition in (5) is different from the one when there are more than two $dist$ left, as depicted in (4). This is because when there are only two $dist$ (3 clusters) left, if the same merging condition as in (4) is used, this may lead to two clusters being merged in which their corresponding mean spectra absorbance distance is slightly less than and nearly equal to the other distance. For example, in Fig. 6, if d_2 is a slightly less than d_1 , then clusters b and c will be merged together. Visibly, it is not convincing. In order to convert this situation to the rest of cases, the merging conditions described in (5) is generated. For example, in Fig. 7, if d_2 is smaller than half of d_1 (similar to there are three $dist$ case) or the extra distance of d_2 to the half of d_1 is less than one tenth of d_2 , then cluster b and c are merged together.

Entire automated cluster merging procedure

In summary the whole procedure of the automated merging clustering algorithm can be described in the following steps.

- (1) The FTIR spectra are initially pre-processed to account for irregularities in cell density across the tissue section. This includes baseline correction and peak area normalization.
- (2) PCA is applied to the processed dataset to reduce its dimensionality. Only the first 10 PCs are extracted and utilized for subsequent FCM cluster analysis.
- (3) The FCM based clustering algorithm is applied to this reduced dataset and optimal clustering structure is adopted by finding the best clustering C_{best} with the minimal V_{SWJ} value.
- (4) The merge cluster algorithm then identifies the reference frequency at which the variance is maximal for the calculated mean spectra.
- (5) The algorithm merges the calculated clusters until a stop criterion reaches.
- (6) The final clustering results is correlated to the original 2D-FTIR image that was collected. Each pixel/spectrum within this image is designated a colour dependent on its cluster membership.

To demonstrate the new algorithm in its entirety, we again utilize the same dataset. The clustering results are shown in Fig. 8. This can be compared to Fig. 1 and demonstrates the successful merging of spectra from regions of cancerous tissue.

7 Experiments and results

7.1 Tissue sample

Lymph node tissue specimens were collected from three different patients during routine surgical resection for breast cancer. Small portions of one axillary lymph node were collected from each case and immediately snap frozen in liquid nitrogen to maintain their biochemical condition. Specimens were subsequently cut into 7 μm thick tissue sections by use of a freezing microtome, placed onto barium fluoride disc, and stored in a cryovial ready for infrared analysis. The remainder of each sample was then fixed in formalin and wax embedded in the traditional manner. This allowed an adjacent tissue section to be cut and further hemotoxylin and eosin (H&E) stained for comparative histological analysis by a Consultant Breast Histopathologist. Two of the tissue sections analyzed were diagnosed as being malign or cancerous in nature, and featured several regions of cancerous invasion. In contrast, the third tissue section analyzed was diagnosed as being benign or healthy in nature. Currently, the

clinical analysis results from the histopathologist are considered as ‘gold standard’ in our experimental analysis.

7.2 Mid-infrared imaging

Prior to infrared data collection, tissue sections were removed from the cryovial and allowed to passively warm to room temperature. By use of the adjacent H&E stained sections, areas of interest upon the sample were first located and then infrared images were obtained using the commercially available Perkin Elmer Spectrum Spotlight 300 FTIR Imaging System[®]. The imaging system utilizes a bi-linear array of small area detectors that are positioned in a chess board like pattern. By use of an electronic stage, the sample is moved in both the X and Y planes underneath the detector array allowing the creation of a two dimensional array of microscopic IR images. Every pixel contained within the image is representative of one IR spectroscopic measurement. It must be noted that the spatial resolution of the Perkin-Elmer Spotlight Imager when collecting at a pixel size of 6.25 μm is c. 12 \times 12 μm , as determined by the diffraction limit. The instrument is in fact collecting data from a 12 μm spatial area, but is being stepped by 6.25 μm , effectively over sampling by 2 fold. For each IR measurement, 16 spectra were co-added over the range 720–4000 cm^{-1} with a spectral resolution of 8 cm^{-1} . Analysis times varied from thirty minutes to several hours per sample dependent upon their size.

7.3 Data pre-processing

All spectra were uniformly pre-treated before undergoing multivariate analysis. Possible contributions from atmospheric water vapor and CO_2 , which absorb in some regions of the IR spectrum, were removed by an atmospheric correction algorithm integrated into the Perkin Elmer Spotlight Software[®]. Sloping and curved baselines sometimes encountered in spectra, caused by irregularities in cell density across tissue sections, were corrected by applying a 6 base point linear interpolation to all spectra. Baseline points used in this process were located at 4000, 3744, 2200, 1836, 876 and 720 cm^{-1} respectively. Finally, spectra were scaled before cluster analysis such that the sum over the indicated wavelengths (720–4000 cm^{-1}) equals unity (also known as peak area normalization).

7.4 Clustering results

In these experiments, three lymph node tissue sections were examined using our newly developed algorithm. The collected IR spectral datasets from each tissue section were initially clustered using the FCM based clustering algorithm. The generated clusters were then combined using our newly

Fig. 8 The extracted spectral dataset after applying the proposed automated merging cluster method

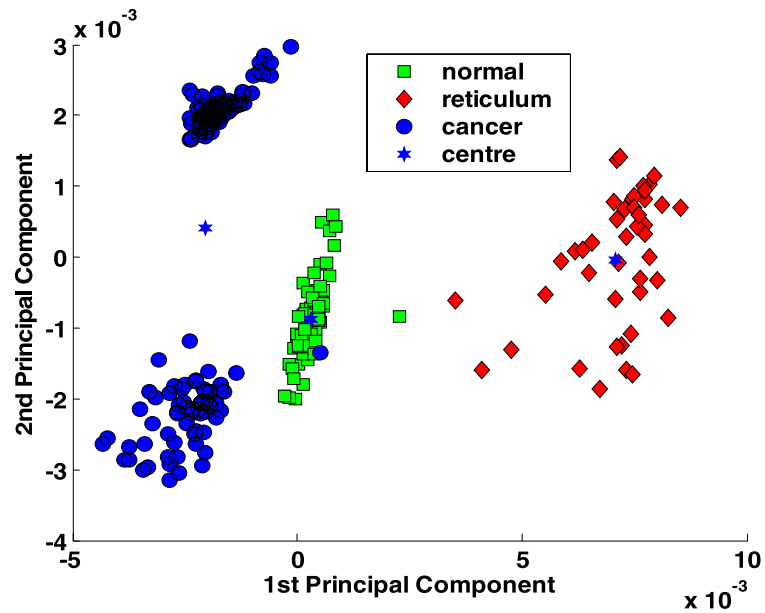
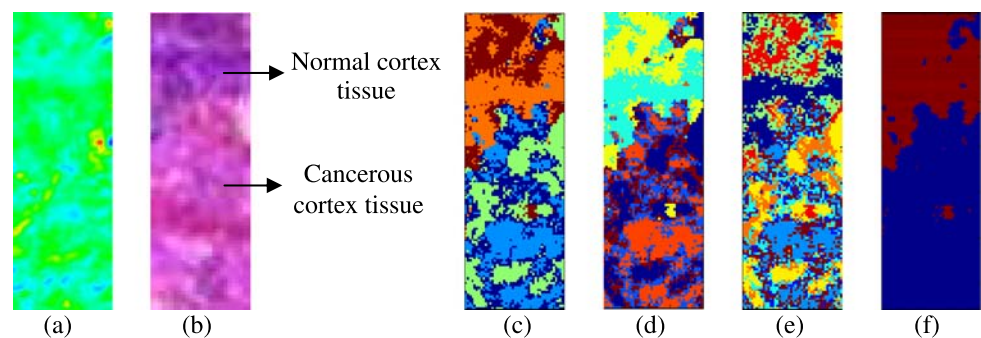


Fig. 9 Lymph node tissue section LNII7. Sampled area was $275 \mu\text{m} \times 818.75 \mu\text{m}$ in size. **a** Total absorbance IR image. **b** H&E stained image. Clustering results after FCM based clustering algorithm. Each colour represents a different cluster of IR spectra. **c** 5 cluster image. **d** 6 cluster image. **e** 9 cluster image. **f** Final results obtained from automated merge clustering algorithm—this image contained two final clusters of IR spectra



proposed automated merge clustering method to help define the main types of tissue found in each section. Due to the effect of randomized initialization, final clustering results may vary. Therefore, we ran the FCM based clustering algorithm ten times for each dataset. Results are displayed in Figs. 8–10.

The tissue section displayed in Fig. 9 was collected from a positive axillary lymph node named LNII7 that displayed large areas of malignancy. Invading cancer from the breast had almost fully infiltrated the lymph node with only small remnants of normal nodal tissue remaining. Figure 9(a) displays the total absorbance IR image collected from an area on the tissue section where both cancerous and normal nodal tissue existed. The IR spectral dataset comprised of 5764 spectra. Figure 9(b) displays the H&E stained image for the same area collected from the parallel stained tissue section. Initial clustering results from the FCM based selection algorithm are shown in Figs. 9(c–e) and the final merged cluster image is shown in Fig. 9(f).

The second tissue section was also taken from a positive axillary lymph node named LNII5. The area studied by IR now displayed several different types of tissue and the existence of a secondary follicle that comprised of proliferating B-lymphocytes. Both the total absorbance IR image and the H&E stained image are shown in Figs. 10(a) and 10(b) respectively. The IR spectral dataset for the examined region comprised a total of 7497 spectra. Only two clustering results were output from the initial clustering algorithm, and are displayed in Figs. 10(c) and 10(e). The corresponding merged cluster image from both clustering structures resulted in three final clusters, as shown in Figs. 10(d) and 10(f).

The final tissue section examined, named LN57, was collected from a benign axillary lymph node. This node had been surrounded by large areas of fatty tissue that had in some regions infiltrated close to the capsule of the node. An infrared image was collected from this region which comprised 7216 spectra. The total IR absorbance and H&E stained images from the examined sample area are shown in

Fig. 10 Lymph node tissue section LNII5. Sampled area was $30625 \mu\text{m} \times 95625 \mu\text{m}$ in size. **a** Total absorbance IR image. **b** H&E stained image. Clustering results after FCM based clustering algorithm. Each colour represents a different cluster of IR spectra. **c** 5 cluster image. **d** Merged cluster result from 5 cluster image. **e** 4 cluster image. **f** Merged cluster result from 4 cluster image. Both merged cluster results contained three clusters of IR spectra

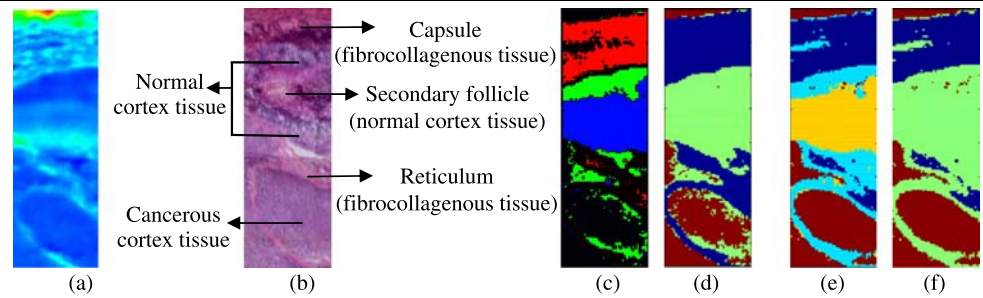
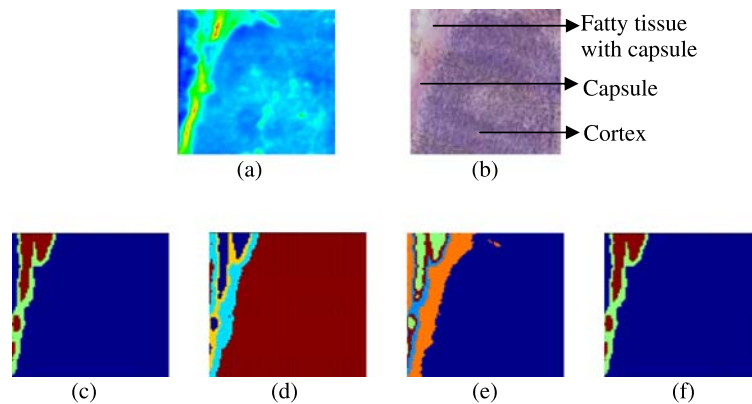


Fig. 11 Lymph node tissue section LN57. Sampled area was $550 \mu\text{m} \times 512.5 \mu\text{m}$ in size. **a** Total absorbance IR image. **b** H&E stained image. Clustering results after FCM based clustering algorithm. Each colour represents a different cluster of IR spectra. **c** 3 cluster image. **d** 4 cluster image. **e** 5 cluster image. **f** Final result obtained from automated merge clustering algorithm. Image contained three final clusters of IR spectra



Figs. 11(a–b). Initial clustering produced three different results, as shown in Figs. 11(c–e). Further cluster merging resulted in all images being made up of three clusters as shown in Fig. 11(f).

To help understand these results from the initial FCM based clustering algorithm, clustering results obtained have been listed in Table 1. This gives the number of clusters that were initially determined by the FCM based algorithm, and the number of clusters they were finally merged into. The value enclosed in parentheses indicates the number of times that specified number of clusters was returned by the FCM based clustering algorithm. For example, when studying the results for lymph node LNII7, the initial algorithm obtained five clusters in two of the runs, six clusters in five of the runs, and nine clusters in the remaining three runs.

8 Discussion

When scrutinizing the results from the initial FCM based clustering algorithm, a number of different cluster results were obtained. This may be due to the random initialization at the beginning of the clustering process. A suitable initial cluster configuration may prevent the occurrence of this problem; however, this can be considered as a research topic in its own right.

Table 1 The number of clusters obtained after the initial FCM based clustering and after cluster merging for repeated runs

	Number of clusters after	
	FCM based clustering algorithm	automated merging clustering method
LNII7	5(2)	2
	6(5)	2
	9(3)	2
LNII5	5(9)	3
	4(1)	3
LN57	3(7)	3
	4(1)	3
	5(2)	3

A tendency to produce an excessive number of clusters was found when applying the FCM based clustering algorithm, especially so for tissue section LNII7. The additional clusters created by the algorithm may describe potential subtypes of tissue that are presently not identified by conventional histopathology. After the application of the novel cluster merging algorithm, a more stable clustering result was obtained. These more clearly described the main types of tissue that existed within the samples analyzed.

As shown in the H&E stained image for tissue section LNII7 (Fig. 9(b)), the invading cancerous tissue exists at the bottom of the collected sample area (pink coloration) and the normal tissue at the top (purple coloration). When studying the initial clustering results (Figs. 9(c–e)), it can be seen that several extra clusters have been created in both the normal and cancerous areas. These clusters found in the area diagnosed as being cancerous could be representative of several different sub-classes of malignancy not normally recognized by histology. In contrast, the extra clusters found in the normal area could be descriptive of normal tissue that is beginning to take on cancerous characteristics. However, at this stage of study, it would be preferable to merge possible subtypes of tissue into one defining group. This would allow a simplified characterization of the tissue section to be made. After applying the newly proposed merge method, the initial overly complex images have now been merged into a single more simplified one (Fig. 9(f)). This newly created image is now more representative of the main characteristics of the tissue section.

The area examined on tissue section LNII5 revealed the more complex infrastructure of a lymph node (Fig. 10(b)). Tissue types found within the sample area include capsule, reticulum, normal cortex and cancerous cortex. It should also be noted that in the centre of the normal cortex, a small spherical region known as a secondary follicle existed, where rapidly proliferating lymphocytes have congressed. The initial clustering image shown in Fig. 10(c), displaying five clusters of IR spectra, was obtained in nine out of ten repeats. When compared against the H&E stained image for the same area (Fig. 10(b)), the full range of tissue types have been correctly characterized, including the secondary follicle. When applying the automated merging method, the clusters which are merged have similar biochemistry, producing three main clusters that describe the main types of tissue found within the sample area (Fig. 10(d)). These include fibrocollagenous tissue (capsule and reticulum), normal cortex tissue (normal cortex and secondary follicle) and finally cancerous cortex tissue.

The second and rare type of output from initial clustering is shown in Fig. 10(e), in which four clusters were obtained for the same IR spectral dataset. The main difference between this output and the first is the incorrect grouping of both reticulum and normal cortex into the same cluster. This is a consequence of the spectra for these types of tissue being close in proximity to each other in PCA space. Due to random initialization of cluster centres, the algorithm has on this occasion calculated that four clusters would best describe the data structure, having the minimum validity index value. Hence these two types of tissue spectra were not separated but grouped into the same cluster. Consequently, after the merging process, reticulum was now grouped with normal cortex tissue rather than capsule tissue. However, the

merge method still discriminated capsule, normal cortex and cancerous cortex tissue.

Tissue section LN57 was collected from a benign lymph node that exhibited large surrounding areas of fatty tissue. The sample area analyzed via IR contained three main types of tissue. These were normal cortex tissue (mainly made up of a secondary follicle), capsule tissue, and small pockets of fatty tissue that had in some regions infiltrated the capsule of the node. Three results were output from the initial clustering algorithm as shown in Figs. 11(c–e). The three clusters shown in Fig. 11(c) characterize the main tissue types found in the examined sample area. The green colour in the image describes the capsule of the lymph node, whereas the red areas represent locations of fatty tissue invasion. Finally the blue colour is descriptive of normal cortex tissue.

The second type of output, as shown in Fig. 11(d), displays four clusters of tissue spectra. Again the capsule has been separated into two clusters that describe areas with or without fatty tissue (blue and yellow colors respectively). But on this occasion an area can be seen that lies beneath the capsule of the lymph node (cyan colour). This is likely to describe a region of the cortex called the subcapsular sinus that lies directly underneath the capsule and allows lymph to enter the node. The final type of output comprised five clusters, as displayed in Fig. 11(e). A similar scenario to the previous has occurred. However, on this occasion an additional cluster has been created that may further describe a layer of fatty tissue within the capsule (red colour). After the automated cluster merging method was applied to these different outputs, a final result comprising three clusters was obtained, corresponding to the three tissue types present. It should be noted that when the initial output was for three clusters, the merge method did not attempt to further combine these, verifying the robustness of this cluster structure.

From these experiments, it can be seen that the proposed automated cluster merging method can rapidly and efficiently obtain major types of biochemical tissue from the existing samples. However, in order to transfer this algorithm into a clinical setting an extensive verification and evaluation process will need to be carried out.

9 Conclusion

FTIR spectroscopic imaging is rapidly becoming a powerful tool in the chemical characterization and monitoring of biological systems. Recent advances in instrumentation have allowed the rapid acquisition of spectra from large spatial areas, such as an entire lymph node. This can allow the construction of chemical images, similar to that of conventional staining that discriminate between different tissue types due to their individual chemical compositions. Multivariate techniques have often been used to analyze the large and com-

plex spectral datasets that are produced. A number of different clustering methods have previously shown an ability to discriminate between tissue pathologies. However, a problem in such unsupervised methods is that the number of clusters that best describe a dataset must be specified in advance. In this contribution, we apply an FCM based clustering algorithm to automatically generate the ‘optimal’ data structure (that giving the minimum validity index value). However, due to the complexity of biological systems an excessive number of clusters can often be obtained. In order to address this problem, an automated cluster merging algorithm has been developed. To demonstrate this proposed algorithm, three axillary lymph node tissue sections have been analyzed using this method. Results indicate the successful merging of clusters that have similar biochemistry, thus demonstrating that the algorithm can successfully determine the main tissue types within the sections.

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.

References

- Johnson KS, Chicken DW, Pickard DCO, Lee AC, Briggs G, Falzon M, Bigio IJ, Keshtgar MR, Bown SG (2004) Elastic scattering spectroscopy for intraoperative determination of sentinel lymph node status in the breast. *J Biomed Opt* 9:1122–1128
- Godavarty A, Thompson AB, Roy R, Eppstein MJ, Zhang C, Gurfinkel M, Sevcik-Muraca EM (2004) Diagnostic imaging of breast cancer using fluorescence-enhanced optical tomography: phantom studies. *J Biomed Opt* 9:486–496
- Fernandez DC, Bhargava R, Hewitt SM, Levin IW (2005) Infrared spectroscopic imaging for histopathologic recognition. *Nat Biotechnol* 23:469–474
- Lasch P, Haensch W, Lewis EN, Kidder LH, Naumann D (2002) Characterization of colorectal adenocarcinoma sections by spatially resolved FT-IR microspectroscopy. *Appl Spectr* 56:1–9
- Chiriboga L, Xie P, Yee H, Zarou D, Zakim D, Diem M (1998) Infrared spectroscopy of human tissue. IV: detection of dysplastic and neoplastic changes of human cervical tissue via infrared microscopy. *Cell Mol Biol* 44:219–230
- Choo LP, Wetzel DL, Halliday WC, Jackson M, LeVine SM, Mantsch HH (1996) In situ characterization of beta-amyloid in Alzheimer’s diseased tissue by synchrotron Fourier transform infrared microspectroscopy. *Biophys J* 71:1672–1679
- Wood BR, Chiriboga L, Yee H, Quinn MA, McNaughton D, Diem M (2004) Fourier transform infrared (FTIR) spectral mapping of the cervical transformation zone, and dysplastic squamous epithelium. *Gynecol Oncol* 93:59–68
- Romeo MJ, Diem M (2005) Infrared spectral imaging of lymph nodes: strategies for analysis and artifact reduction. *Vib Spectr* 38:115–119
- Lasch P, Haensch W, Naumann D, Diem M (2004) Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis. *Biochim Biophys Acta Mol Basis Dis* 1688:176–186
- Wang XY, Whitwell G, 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 systems design and applications*, Budapest, Hungary, pp 467–472
- Wang XY, Garibaldi JM (2005) Simulated annealing fuzzy clustering in cancer diagnosis. *Eur J Inf* 29:61–70
- Wang XY, Garibaldi JM, Bird B, George MW (2005) Fuzzy clustering in biochemical analysis of cancer cells. In: *Proceedings of the fourth conference of the European society for fuzzy logic and technologies (EUSSFLAT 2005)*, Barcelona, Spain, pp 1118–1123
- Sun H, Wang S, Jiang Q (2004) FCM-based model selection algorithms for determining the number of clusters. *Pattern Recognit* 37:2027–2037
- Jolliffe IT (1986) *Principal component analysis*. Springer, New York
- Ward JH (1963) Hierarchical grouping to optimize an objective function. *J Am Stat Assoc* 58(301):236–244
- Xie Y, Raghavan VV, Zhao X (2002) 3M algorithm: finding an optimal fuzzy cluster scheme for proximity data. In: *Proceedings of the FUZZY—IEEE conference—2002 IEEE world congress on computational intelligence*, Honolulu, HI, vol 1, pp 627–632
- Richter T, Steiner G, Abu-Id M, Salzer R, Bergmann R, Rodig H, Johannsen B (2002) Identification of tumour tissue by FTIR spectroscopy in combination with positron emission tomography. *Vib Spectr* 28:103–110
- Stone N, Kendall C, Shepherd N, Crow P, Barr H (2002) Near-infrared Raman spectroscopy for the classification of epithelial pre-cancers and cancers. *J Raman Spectr* 33:564–573
- Lasch P, Wasche W, McCarthy WJ, Muller G, Naumann D (1998) Imaging of human colon carcinoma thin sections by FTIR microspectrometry. *SPIE Proc* 3257(3):187–197
- Kim SW, Ban SH, Chung H, Cho S, Chung HJ, Choi PS, Yoo OJ, Liu JR (2004) Taxonomic discrimination of flowering plants by multivariate analysis of Fourier transform infrared spectroscopy data. *Plant Cell Rep* 23(4):246–250
- Liu H, Motoda H (1998) *Feature selection for knowledge discovery and data mining*. Kluwer Academic, Boston
- Bezdek J (1981) *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, New York
- Bezdek J (1998) *Pattern Recognition*. In: Ruspini, EH, Bonissone PP, Pedrycz W (eds) *Handbook of fuzzy computation*. IOP, Boston
- Dunn JC (1973) A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J Cybern* 3(3):32–57
- Rezaee MR, Lelieveldt BPF, ReiBer JHC (1998) A new cluster validity index for the fuzzy c-means. *Pattern Recognit Lett* 19:237–246
- Ray S, Turi R (1999) Determination of number of clusters in k-means clustering and application in colour image segmentation. In: *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*, pp 137–143
- Xu L (1996) How many clusters? A Ying–Yang machine based theory for a classical open problem in pattern recognition. In: *Proceedings of 1996 IEEE international conference on neural networks*, pp 1546–1551
- Berkhin P (2002) *Survey of clustering data mining techniques*. Technique Report, Accrue Software, San Jose, CA
- Xu R, Wunsch II D (2005) Survey of clustering algorithm. *IEEE Trans Neural Netw* 16(3):645–678
- Ball G, Hall D (1965) ISODATA, a novel method of data analysis and classification. Technical Report AD-699616, Stanford University, Stanford, CA
- Turi RH (2001) *Clustering-based colour image segmentation*. PhD thesis, School of Computer Science and Software Engineering, Monash University, Australia

32. Chaudhuri D, Chaudhuri BB, Murthy CA (1992) A new split-and-merge clustering technique. *Pattern Recognit Lett* 13:399–409
33. Tou JT (1979) DYNOC—A dynamic optimal cluster-seeking technique. *Int J Parallel Prog* 8(6):541–547
34. Tseng L, Yang S (2001) A genetic approach to the automatic clustering problem. *Pattern Recognit* 34:415–424
35. Garai G, Chaudhuri BB (2004) A novel genetic algorithm for automatic clustering. *Pattern Recognit Lett* 25:173–187
36. Zhang DQ, Chen SC (2004) A novel kernelized fuzzy c-means algorithm with application in medical image segmentation. *Artif Intell Med* 32:37–50
37. Chen SC, Zhang DQ (2004) Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure. *IEEE Trans Syst Man Cybern Part B Cybern* 34(4):1907–1916