Simulated Annealing Fuzzy Clustering in Cancer Diagnosis

Xiao-Ying Wang and Jonathan M. Garibaldi
Automated Scheduling, Optimisation and Planning (ASAP) Research Group
Department of Computer Science & Information Technology
The University of Nottingham, Jubilee Campus, Wollaton Road, United Kingdom
{xyw, jmg}@cs.nott.ac.uk

Keywords: Fourier Transform Infrared spectroscopy, Hierarchical Cluster Analysis, Fuzzy C-Means, Simulated Annealing Fuzzy Clustering, Xie-Beni validity measure.

Received: November 10, 2004

Classification is an important research area in cancer diagnosis. Fuzzy C-means (FCM) is one of the most widely used fuzzy clustering algorithms in real world applications. However there are two major limitations that exist in this method. The first is that a predefined number of clusters must be given in advance. The second is that the FCM technique can get stuck in sub-optimal solutions. In order to overcome these two limitations, Bandyopadhyay proposed a Variable String Length Simulated Annealing (VFC-SA) algorithm. Nevertheless, when this algorithm was implemented, it was found that sub-optimal solutions were still obtained in certain circumstances. In this paper, we propose an alternative fuzzy clustering algorithm, Simulated Annealing Fuzzy Clustering (SAFC), that improves and extends the ideas present in VFC-SA. The data from seven oral cancer patients tissue samples, obtained through Fourier Transform Infrared Spectroscopy (FTIR), were clustered using FCM, VFC-SA and the proposed SAFC algorithm. Experimental results are provided and comparisons are made to illustrate that the SAFC algorithm is able to find better clusters than the other two methods.

Povzetek: Opisana je nova variacija algoritma FMC za klasifikacijo s pomočjo mehkega grupiranja.

1 Introduction

Cancer has become one of the major causes of mortality around the world and research into its diagnosis and treatment has become an important issue for the scientific community. In Britain, more than one in three people will be diagnosed with cancer during their lifetime and one in four will die from cancer. Accurate diagnostic techniques could enable various cancers to be detected in their infancy and, consequently, the corresponding treatments could be undertaken earlier. In recent years, FTIR has been increasingly applied to the study of biomedical conditions and could become a very powerful tool for determination and monitoring of chemical composition within biological systems [1]. It has also been used as a diagnostic tool for various human cancers and other diseases [2-5]. This technology works by measuring the wavelengths at which different functional groups of chemical samples absorb infrared radiation (IR) and the intensities of these absorptions. The quantity of absorption depends on the chemical bonds and the structure of the molecule and, hence, small changes in molecular structure can significantly affect the absorption intensity. Since chemical functional groups absorb light at specific wavelengths, the resultant FTIR spectrum can be likened to a molecular "fingerprint". If the characteristic spectrum of an abnormal and normal tissue component is known (in a "fingerprint library"), it may be possible to compare each obtained spectrum to these reference spectra and, hence, accurate diagnosis may be achieved. An instance of FTIR spectra from a non-biochemical application in which example spectra for standard and unknown paint samples are compared is shown in Figure 1 [6]. In the context of cancer diagnosis, the FTIR technique detects molecular differences within the cell rather than morphological changes of the cell and hence may lead to earlier detection of cell abnormalities.

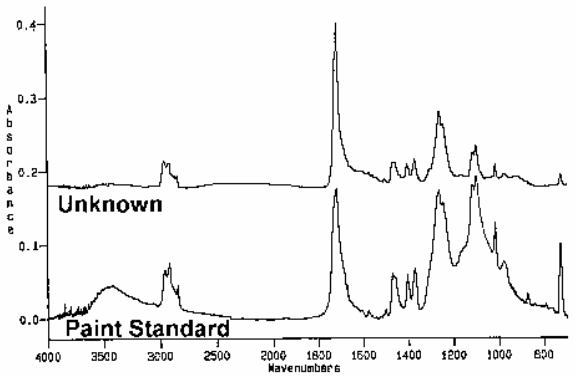

Figure 1. FTIR spectra for paint analysis.

Some advantages of FTIR analysis compared to conventional cytological clinical analysis might be:

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 so 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 previous clinical work [7], Chalmers *et al.* reported on the analysis of sets of FTIR spectra taken from oral cancer tissue samples. In general, the experiments analyzed the tissue samples in two parallel processes. In the first process, the samples were scanned by FTIR spectroscopy, various pre-processing techniques (such as mean-centering, variance scaling and first derivative) were performed on the FTIR spectral data empirically. The data was then classified by hierarchical cluster analysis after principal component analysis. In the second process, the samples were stained with a chemical solution and then examined through conventional cytology to group the samples into different functional groups. The results from these two processes were then compared. The clustering results 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. However, the pre-processing procedures needed extra time, software tools and significant human expertise. If a clustering technique could be developed which could obtain clustering results as good or even better than conventional clinical analysis without the necessity for pre-processing procedures, it would make the diagnosis more efficient and enable automation.

In previous research work, hierarchical clustering analysis (HCA) and the fuzzy c-means (FCM) algorithm have been used to classify non pre-processed FTIR oral cancer data [8]. The results showed that the FCM method performed significantly better than HCA. However, there are two major limitations of FCM which may affect the use of the technique as a practical diagnostic tool. Firstly, before performing the algorithm, an assumption of the number of clusters has to be made in advance. In real medical diagnosis, of course this number would not be known. Secondly, it is a non-convex method [9] so may often lead to local minima solutions, and hence misdiagnosis could occur. In order to avoid these limitations, a simulated annealing based FCM algorithm (SAFC) was introduced by the authors in [10]. It was developed by modifying and extending Bandyopadhyay's Variable String Length Simulated Annealing (VFC-SA) algorithm which was used for the classification of remote sensing satellite images [11].

In this paper, we describe the SAFC algorithm in further detail and give additional analysis on the experimental results. In Section 2, the background techniques and some related work are introduced. The original VFC-SA and our extended SAFC algorithm are described in Section 3. In Section 4, we provide the results of our experimentation in which the FCM, VFC-SA and SAFC algorithms were applied to seven sets of oral cancer FTIR data. The classification results are discussed in Section 5 and conclusions are drawn.

2 Background

2.1 FCM algorithm

The FCM algorithm, also known as Fuzzy ISODATA, is one of the most frequently used methods in pattern recognition. It is based on minimisation of the objective function (1) to achieve good classifications.

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

J(U,V) is a squared error clustering criterion, and solutions of minimisation of (1) are least-squared error stationary points of *J(U,V)*. The expression, $X = \{x_1, x_2, ...x_n\}$ is a collection of data, where *n* is the number of data points. $V = \{v_1, v_2, ...v_c\}$ is a set of corresponding cluster centres in the data set ***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:

$$\mu_{ij} \in [0,1], \qquad \forall i = 1,...n, \forall j = 1,...c \qquad \textbf{(2)}$$

$$\sum_{j=1}^{c}\mu_{ij} = 1, \qquad \forall i = 1,...n \qquad \textbf{(3)}$$

Where $U = (\mu_{ij})_{n*c}$ is a fuzzy partition matrix, $\| x_i - v_j \|$ represents the Euclidean distance between x_i and v_j, parameter *m* is the "fuzziness index" and is used to control the fuzziness of membership of each datum in the range $m \in [1, \infty]$. In this experimentation the value of $m = 2.0$ was chosen. Although there is no theoretical basis for the optimal selection of *m*, this has been chosen because the value has been commonly applied within the literature. The FCM algorithm is described in, for example, [12] and can be

performed by the following steps:

1) Initialize the cluster centres $V = \{v_1, v_2, ...v_c\}$, or initialize the membership matrix μ_{ij} with random value and make sure it satisfies conditions **(2)** and **(3)** and then calculate the centres.

2) Calculate the fuzzy membership μ_{ij} using

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} (\frac{d_{ij}}{d_{ik}})^{\frac{2}{m-1}}} \quad \textbf{(4)}$$

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

3) Compute the fuzzy centres v_j using

$$v_j = \frac{\sum_{i=1}^{n} (\mu_{ij})^m x_i}{\sum_{i=1}^{n} (\mu_{ij})^m}, \forall j = 1,...,c \quad \textbf{(5)}$$

4) Repeat steps 2) and 3) until the minimum *J* value is achieved.

5) Finally, defuzzification is necessary to assign each data point to a specific cluster (i.e. by setting a data point to a cluster for which the degree of the membership is maximal).

2.2 Simulated Annealing algorithm and related works

The first simulated annealing algorithm was proposed by Metropolis et al. in 1953 [13]. It was motivated by simulating the physical process of annealing solids. The process can be described as follows. Firstly, a solid is heated from a high temperature and then cooled slowly so that the system at any time is approximately in thermodynamic equilibrium. At equilibrium, there may be many configurations with each one corresponding to a specific energy level. The chance of accepting a change from the current configuration to a new configuration is related to the difference in energy between the two states. Kirkpatrick et al. were the first to introduce simulated annealing to optimisation problems in 1982 [14]. Since then, simulated annealing has been widely used in combinatorial optimisation problems and has achieved good results on a variety of problem instances.

We use E_n and E_c represent the new energy and current energy respectively. E_n is always accepted if it satisfies $E_n < E_c$, but if $E_n \geq E_c$ the new energy level is only accepted with a probability as specified by $\exp(-(E_n - E_c)/T)$, where T is the current temperature. Hence, worse solutions are accepted based on the change in solution quality which allows the search to avoid becoming trapped at local minima. The temperature is then decreased gradually and the annealing process is repeated until no more improvement is reached or any termination criteria have been met.

Al-Sultan [15,16] and Kein and Dubes [17] have developed algorithms based on simulated annealing to find the global minimum solution using Fuzzy C-Means and other crisp (non-fuzzy) clustering methods. These were applied, for example, to determine the best clustering criterion for the multi-sensor fusion problem. However, the number of clusters has to be declared in advance for both of these techniques.

Although simulated annealing is used in the experimentation described here, other search algorithms have been used by other authors. Tseng and Yang proposed a genetic algorithm based clustering algorithm, in which the genetic algorithm was used to group the small clusters into successively larger clusters. A heuristic strategy [18] is then used to find a 'good' clustering (see below). Maulik and Bandyopadhyay developed a fuzzy clustering method which combined a genetic algorithm and FCM clustering to automatically segment satellite images obtained by remote sensing [19].

2.3 Xie-Beni validity index

Clustering validity is a concept that is used to evaluate the quality of clustering results. If the number of clusters is not known prior to commencing an algorithm, the clustering validity index may be used to find the optimal number of clusters [20]. This can be achieved by evaluating all of the possible clusters with the validity index and then the optimal number of clusters can be determined by selecting the minimum value of the index.

Many clusters validation indices have been developed in the past. In the context of fuzzy methods, some of them only use the membership values of a fuzzy cluster of the data, such as the partition coefficient [21] and partition entropy [22]. The advantage of this type of index is that it is easy to compute but it is only useful for the small number of well-separated clusters. Furthermore, it also lacks direct connection to the geometrical properties of the data. In order to overcome this problem Xie and Beni defined a validity index which measures the compactness and separation of clusters [23]. In this paper, the Xie-Beni index has been chosen as the cluster validity measure because it has been shown to be able to detect the correct number of clusters in several experiments [24]. Xie-Beni validity is the combination of two functions. The first calculates the compactness of data in the same cluster and the second computes the separateness of data in different clusters. Let S represent the overall validity index, π be the

compactness and s be the separation of the fuzzy c-partition of the data set. The Xie-Beni validity can now be expressed as:

$$S = \frac{\pi}{s} \qquad \textbf{(6)}$$

where

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

and

$$s = (d_{\min})^2.$$

$d_{\min}$ is the minimum distance between cluster centres, given by $d_{\min} = \min_{ij} \| v_i - v_j \|$.

Smaller values of π indicate that the clusters are more compact and larger values of s indicate the clusters are well separated. Thus a smaller S reflects that the clusters have greater separation from each other and are more compact.

3 VFC-SA and SAFC

Recently, Bandyopadhyay proposed a Variable String Length Simulated Annealing (VFC-SA) algorithm [11]. It 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 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.

3.1 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:

$$1/(1+\exp(-(E_n - E_c)/T))$$

The Xie-Beni index was used to compute the evaluation of a cluster. 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) was used following the suggestion proposed by Bezkek in [25]. The initial temperature T was set to a high temperature $T_{\max}$, a neighbour of the solution was produced by randomly flipping one bit within the string (describing the cluster centres) 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 $T = rT$ and was repeated until the 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:

Set parameters $T_{\max}, T_{\min}, c, k, r$.

Initialised the string by randomly choosing c data points from the data set to be cluster centres.

Compute the corresponding membership values using equation **(4)**

Calculate the initial energy E_c using XB index from equation **(6).**

Set the current temperature $T = T_{\max}$.

while $T \geq T_{\min}$

For $i = 1$ to k

Perturb a current centre in the string.

Compute the corresponding membership values using equation **(4)**.

Compute the corresponding centres with the equation **(5)**.

Calculate the new energy E_n from the new string.

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.

Else, reject it.

End for

$T = rT$.

End while.

Return the current string as the final solution.

The process of perturbing a current cluster centre comprised three functions. They are: perturbing an existing centre (*Perturb Centre*), splitting an existing

centre (*Split Centre*) and deleting an existing centre (*Delete Centre*). At each iteration, one of the three functions was randomly chosen. When splitting or deleting a centre, the cluster sizes were used to select a centre. The size, C_j, of a cluster, j, can be expressed by (where c is the number of clusters):

$$| C_j | = \sum_{i=1}^{n} \mu_{ij} , \quad \forall j = 1,...c \qquad \textbf{(7)}$$

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 $d = 1,...,N$, 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. Let $v_{current}[d]$ and $v_{new}[d]$ represent the current and new centre respectively, and *Perturb Centre* can then be expressed as:

$$v_{new}[d] = v_{current}[d] + cr[d] .$$

b) *Split Centre*

The centre of the biggest cluster is chosen by using equation **(7)**. This centre is then replaced by two new centres which are created by the following procedure. A reference point with a membership value less than but closest 0.5 to the selected centre is identified. Then the distance between this reference point and the current chosen centre is calculated using:

$$dist[d] = | v_{current}[d] - w_{reference}[d] |$$

Finally, the two new centres are then obtained by:

$$v_{new}[d] = v_{current}[d] \pm dist[d]$$

c) *Delete Centre*

As opposed to *Split Centre*, the smallest cluster is identified and its centre deleted from the string encoding.

3.2 SAFC algorithm

When the original VFC-SA algorithm was implemented by the authors on a wider set of test cases than originally used by Bandyopadhyay [11], it was found to suffer from several difficulties. In order to overcome these difficulties, four extensions to the algorithm were developed. In addition, some details were not explicit in the original algorithm. 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. This is because re-initialization is a source of computational inefficiency. Using the clustering results from previous results leads to a better initialization.

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 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 closest 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 2 (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 less than 0.5 to v_1). 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 would 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 2 is used to illustrate this process. A data point is chosen, w_1, that is closest the mean value of the membership degree 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 3. Obviously the new centres are better than the ones in Figure 2 and therefore better solutions are likely to be found in same time (number of iterations).

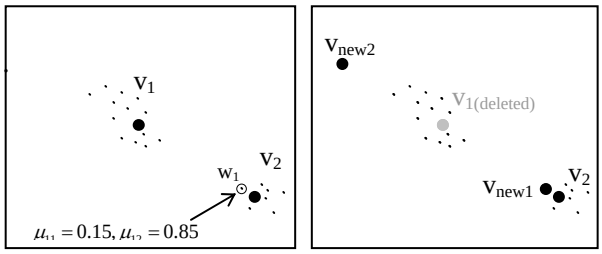

Figure 2. An illustration of *Split Centre* from the original algorithm with distinct clusters (where μ_{11} and μ_{12} represent the membership degree of w_1 to the centres v_1 and v_2 respectively)

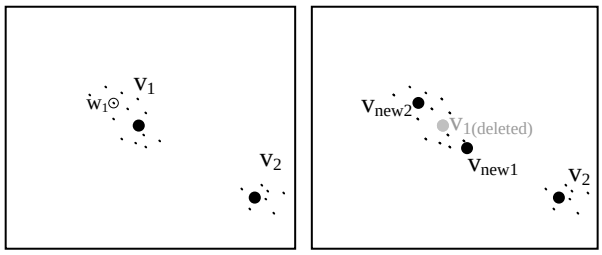

Figure 3. The new *Split Centre* applied to the same data set as Figure 2, above, (where w_1 is now the data point that is closest to the mean value of the membership degree above 0.5)

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 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 never violates the criteria whereby the number of clusters C should be within the range of $[c_{\min}, c_{\max}]$. Therefore when splitting a centre, if the number of clusters has reached $c_{\max}$ then the operation is disallowed. Dually, when deleting a centre, the operation is not allowed if the number of clusters in the current solution is $c_{\min}$.

4 Experiments and Results

In this section, the clinical data used are firstly introduced and then the FCM, VFC-SA and SAFC algorithms are applied to seven sets of oral cancer FTIR data in order to compare the results.

4.1 Clinical data background

In these experiments, all the algorithms are applied to FTIR spectral data sets obtained from oral cancer patients. These data have been provided by Leeds Royal Infirmary, U.K. and Derby Royal Infirmary, U.K. and Derby City General Hospital, UK. All of the FTIR spectra data have been produced by a Nicolet 730 FTIR spectrometer (Nicolet Instruments, Inc., Madison, USA), which is interfaced to a NicPlan IR-microscope fitted with a liquid-nitrogen cooled narrow-band mercury-cadmium-telluride (MCT) detector. Transmission spectra were recorded either 4cm^{-1} or 8cm^{-1} spectral resolution, typically co-adding 512 or 1024 scans per spectrum. The FTIR microscope was operated using an $32\times$ objective lens. Background single-beam spectra were recorded through a blank BaF2 window. A Nicolet Nexus FTIR spectrometer interfaced to a Continuum IR microscope fitted with a narrow-band MCT detector, sited at the University of Nottingham, was used to record the conventional Globar-sourced spectra.

Multivariate data analysis on pre-processed spectra was undertaken using Infometrix Pirouette, version 3, multivariate analysis software (Infometrix, Inc., Woodinville, WA, USA). In this study, the data analysis was limited to those that lie within the spectral range 900-1800 cm^{-1}

The tissue samples, with nominal thickness $5\mu m$, were mounted on $0.5mm$ thickness BaF_2 windows for FTIR investigations. Parallel sections were stained conventionally to facilitate identifying regions for particular interest. Some of the sections used for infrared examinations were also stained after they had been studied spectroscopically.

In this study, the FCM, VFC-SA and SAFC algorithms were implemented in MATLAB (version 6.5.0, release 13.0.1).

All the FTIR spectra were taken from three oral cancer patients, which contain a mixture of tumour (neoplasm), stroma (connective tissue), ‘early keratinisation’ and ‘necrotic’. The seven data sets have been taken from three different patients. The number of data points within each of these data sets is: 15, 18, 11, 31, 30, 15 and 42. Figure 4 (a) shows a $4\times$ magnification visual image from one of the hematoxylin and Eosin stained oral tissue sections, which has been taken from the first patient. There are two types of cells (stroma and tumour) in this section with their regions clearly identifiable by their light and dark coloured stains respectively. Figure 4(b) shows a $32\times$ 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 using an aperture of $10\mu m \times 10\mu m$. The locations of these are marked by ‘+’ on Figure 4.(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 data set 1, and corresponding FTIR spectra are shown in Figure 5.

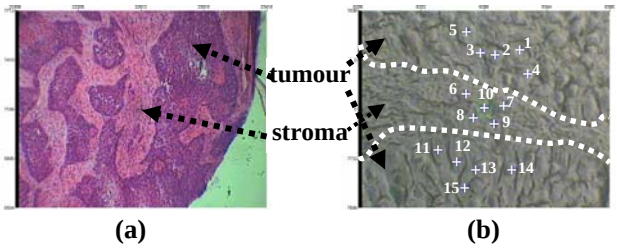

Figure 4. Tissue samples from data set 1 (a) stained (b) unstained

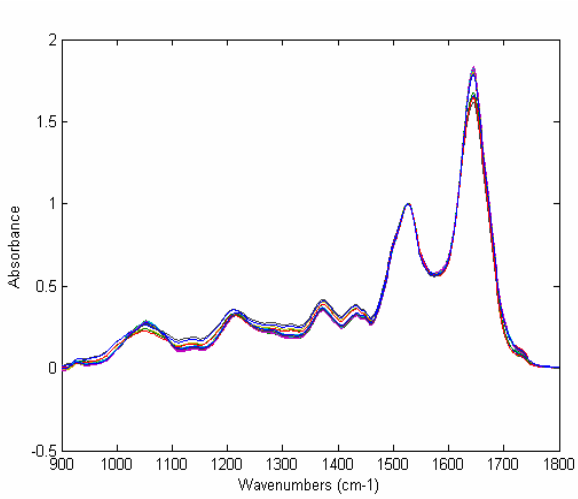

Figure 5. FTIR spectra from data set 1

4.2 Evaluation of FCM, VFC-SA and SAFC

In these experiments, the number of different types of cells in each tissue section from clinical analysis was considered as the number of clusters to be referenced. They were also used as the parameter for FCM. The Xie-Beni index value has been utilised throughout to evaluate the quality of the classification for these three algorithms. The parameters for VFC-SA and SAFC are: $T_{\min} = 1e-5$, $k = 40$, $r = 0.9$. $T_{\max}$ was set as 3 in all cases. That is because the maximum temperature has a direct impact on how much worse the XB index value of 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) [26]. 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. Each data set was allowed 10 runs on each method. 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 XB index value. VFC-SA and SAFC automatically found the number of clusters by choosing the solution with the smallest XB index value. Table 1 shows the average XB index values obtained after 10 runs of each algorithm (best average is shown in bold).

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

Table 1. Average of the XB index values obtained when using the FCM, VFC-SA and SAFC algorithms.

In Table 1, it can be seen that in all of these seven data sets, the average 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 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 XB index values in SAFC is only slightly smaller than that obtained using VFC-SA. Nevertheless, when the Mann-Whitney test (with $p<0.01$) [27] was conducted on the results of these two algorithms, the XB index for SAFC was found to be statistically significantly lower 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 2. The brackets 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.

Dataset	Number of Clusters in Solution		
	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)

Table 2. Comparison of the number of clusters achieved by clinical analysis, VFC-SA and the SAFC methods.

In Table 2, it can be observed that in data sets 3, 4, 5 and 7, either one or both of the 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 even produced a variable number of clusters within the 10 runs. Returning to the XB validity index values of Table 1, it was shown that all the average XB index values obtained by SAFC are better.

It can be observed that the corresponding XB average index values for SAFC for data sets 3, 4 and 5 produced much smaller values than FCM. These three data sets are also the data sets which SAFC obtained a different number of clusters to clinical analysis. In data set 3, the average XB index value in SAFC is much smaller than in VFC-SA. This is because the number of clusters obtained from these two algorithms is different (see Table 2). 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 XB index values are noticeable, though not as big as data set 3. This is because in these two data sets, some runs of VFC-SA obtained the same number of clusters as SAFC.

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 [28, 29]. The data has been plotted in this way because, although the FTIR spectra are limited to within $900cm^{-1} - 1800cm^{-1}$, there are still 901 absorbance values corresponding to each wavenumber 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 visualization of the solutions that have been achieved. Figures 6, 7, 8 and 9 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.

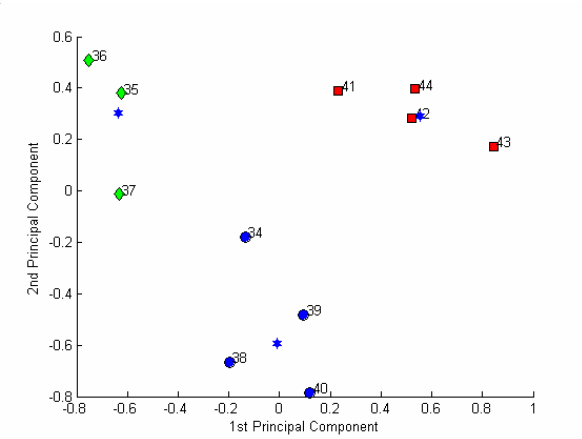

Figure 6. SAFC Cluster result in PCA for data set 3.

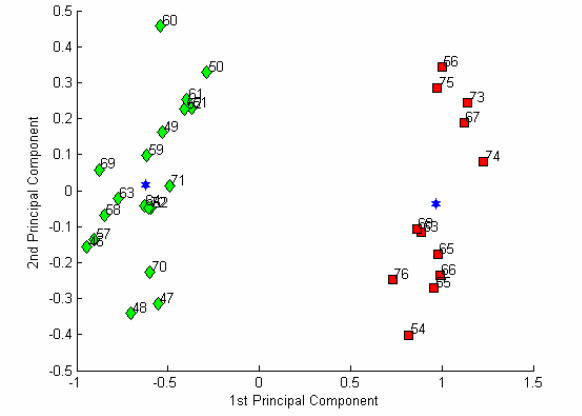

Figure 7. SAFC Cluster result in PCA for data set 4.

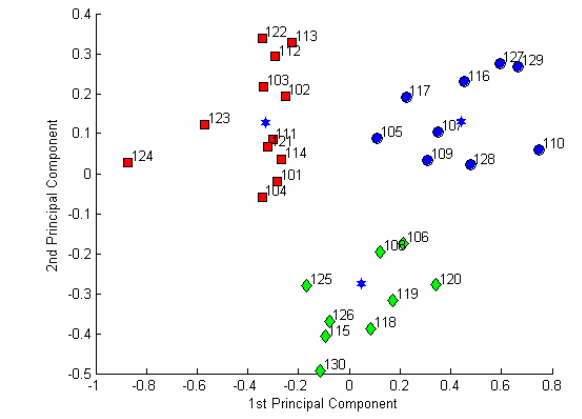

Figure 8. SAFC Cluster result in PCA for data set 5.

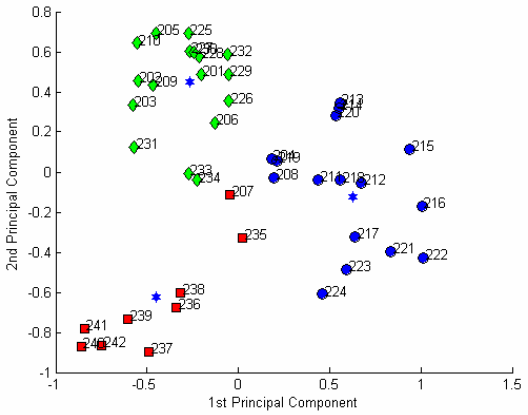

Figure 9. SAFC Cluster result in PCA for data set 7.

There are three possible explanations for this phenomenon. 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 XB validity index value was obtained, indicating a 'better' solution in technical terms, the Xie Beni validity index is not accurately capturing the real validity of the clusters. Put another way, although the SAFC finds the better solution in terms of Xie-Beni validity 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 the SAFC produces better solutions in that it consistently finds better (statistically lower) values of the objective function (Xie-Beni validity index).

5 Conclusion

In this paper, a new SAFC method has been proposed which has been extended from the original VFC-SA algorithm in four ways. The newly proposed algorithm's performance has been evaluated on seven oral cancer FTIR data and compared to clinical analysis, FCM and VFC-SA. The XB validity index was used as the evaluation method to measure the quality of the clusters produced. The experimental results have shown that the SAFC algorithm can escape the sub-optimal solutions obtained in the other two approaches and hence produce better clusters. On the other hand, the number of clusters obtained by SAFC in some data sets are not in agreement with those provided through clinical analysis. This can be explained in three ways. Firstly, the number of cluster from clinical analysis may not correct; secondly, the XB validity index may not suitable to apply on these clinical data; and thirdly, the FTIR technique has not (for these data sets) captured sufficient information to permit correct classification. However, more results and information are needed before any definitive conclusion can be made in this case. Nevertheless, this SAFC algorithm is a further step towards the automatic classification of data for real medical applications. The further development of this algorithm is ongoing research area.

In the future, we are also trying to obtain 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 and lymphnodes disease. Establishing the techniques necessary to develop clinically useful automated diagnosis tools across a range of medical domains is the ultimate goal of this research.

Acknowledgement

The authors are grateful to John Chalmers et al. for providing the FTIR spectral data used in this study, and for making available their internal report on the clinical study carried out at Derby General City Hospital. The authors would like to express their kind thanks to Sanghamitra Bandyopadhyay for permission to use and extend the VFC-SA algorithm. Moreover, the authors also like to express their thanks to the anonymous referees for their valuable comments.

References

[1] H. Mantsch, R. N. McElhaney, "Application of IR spectroscopy to biology and medicine," *J Molec Struc*, vol. 217, pp. 347-362, 1990.

[2] P. T. T. Wong, B. Rigas, "Infrared spectra of microtome sections of human colon tissues," *Applied Spectroscopy*, vol. 44, pp. 1715-1718, 1990.

[3] B. Rigas, S. Morgello, I.S. Goldan, P. T. T. Wong, "Human colorectal cancers display abnormal Fourier –transform infrared spectra," in *Proc. The National Academy of Science of the USA*, vol. 87, 1990, pp. 8140-8144.

[4] P. T. T. Wong, S. M. Goldstein, R. C. Grekin, T. A. Godwin, C. Pivik, B. Rigas, "Distinct infrared spectroscopic patterns of human basal cell carcinoma of the skin," *Cancer Research*, vol. 53, no. 4, pp. 762-765, 1993.

[5] B. J. Morris, C. Lee, B. N. Nightingale, E. Molodysky, L. J. Morris, R. Appio, "Fourier transform infrared spectroscopy of dysplastic, papillomavirus-positive cervicovaginal lavage

speciens," *Gynecological Oncology*, vol. 56., no. 2, pp. 245-249, 1995.

[6] *Handbook of Analytical Method for Materials,* Materials Evaluation and Engineering, Inc, United Stated of America, 2001, pp. 15.

[7] R. Allibone, J. M. Chalmers, M. A. Chesters, S. Fisher, A. Hitchcock, M. Pearson, F. J. M. Rutten, I. Symonds, M. Tobin, "FT-IR microscopy of oral and cervical tissue samples", internal report, Derby City General Hospital, England, 2002, unpublished.

[8] X-Y. Wang, J.M. Garibaldi, and T. Ozen, "Application of The Fuzzy C-Means Clustering Method on the Analysis of non Pre-processed FTIR Data for Cancer Diagnosis", in the Proceedings of the 8th Australian and New Zealand Conference on Intelligent Information Systems, pp. 233-238, Sydney, Australia, December 10-12, 2003.

[9] S. Z. Selim, "Using nonconvex programming techniques in cluster analysis", *Jt Meet. Ops Res. Soc. Am. And Inst. Mgmt Sci*., Houston, Texas, 1981.

[10] X-Y. Wang, G. Whitwell and J.M Garibaldi, "The Application of a Simulated Annealing Fuzzy Clustering Algorithm for Cancer Diagnosis", in the Proceedings of IEEE 4th International Conference on Intelligent Systems Design and Application, pp. 467-472, Budapest, Hungary, August 26-28, 2004.

[11] S. Bandyopadhyay, "Simulated Annealing for Fuzzy Clustering: Variable Representation, Evolution of the Number of Clusters and Remote Sensing Application", Machine Intelligence Unit, Indian Statistical Institute, 2003, unpublished, private communication.

[12] J. Bezdek, Pattern Recognition With Fuzzy Objective Function Algorithms. New York: Plenum, 1981.

[13] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller and E. Teller, "Equation of State Calculations by Fast Computing Machines", *J. Chem. Phys.*, vol. 21, no. 6, pp. 1087-1092, 1953.

[14] S. Kirkpatrick, C. D. Gelatt Jr. and M. P. Vecchi, "Optimisation by Simulated Annealing", *IBM Research Report*, RC 9355, 1982.

[15] K. S. Al-Sultan and S. Z. Selim, "A Global Algorithm for the Fuzzy Clustering Problem", *Pattern Recognition*, vol. 26, no. 9, pp. 1357-1361, 1993.

[16] S. Z. Selim and K. Alsultan, "A Simulated Annealing Algorithm for the Clustering Problem", *Pattern Recognition*, vol. 24, no. 10, pp. 1003-1008, 1991.

[17] R. W. Klein and R. C. Dubes, "Experiments in Projection and Clustering by Simulated Annealing", *Pattern Recognition*, vol. 22. no. 2. pp. 213-220, 1989.

[18] L. Y. Tseng, S. B. Yang, "A genetic approach to the automatic clustering problem", *Patten Recognition*, vol. 34. pp. 415-424, 2001.

[19] U. Maulik and S. Bandyopadhyay, "Fuzzy Partitioning Using a Real-coded Variable Length Genetic Algorithm for Pixel Classification", *IEEE Trans. On Geosciences and Remote Sensing*, vol. 41, no. 5, 5 May 2003.

[20] M. Ramze Rezaee, B. P. F. Leieveldt, J. H. C. Reiber, A New Cluster Validity Index for the Fuzzy C-Means, *Pattern Recognition Letters*, Vol. 19, Elsevier, pp. 237-246, 1998.

[21] J. C.Bezdek, Cluster Validity with Fuzzy Sets. *J. Cybernet*, Vol. 3, No. 3, pp. 58-72, 1974.

[22] J. C. Bezdek, Mathematical Models for Systematics and Taxonomy. In Estabrook, G. (Ed.), Proceeding of 8th Internat Conference Numerical Taxonomy. Freeman, San Francisco, CA, pp. 143-166, 1975.

[23] X. L. Xie and G. Beni, " A validity measure for fuzzy clustering", *IEEE Trans. Pattern Analysis and Machine Intelligence,* Vol. 13, pp. 841-847, 1991.

[24] N. R. Pal, J. C. Bezdek, "On cluster validity for the fuzzy c-means model", *IEEE Trans. Fuzzy Systs.*, Vol. 3, pp. 370-379, 1995.

[25] J. C. Bezdek, *Pattern Recognition* in Handbook of Fuzzy Computation. IOP Publishing Ltdl, Boston, Ny 1998(Chapter F6).

[26] V. J. Rayward-Smith, I. H. Osman, C. R. Reeves and G. D. Smith, Modern Heuristic Search Methods. John Wiley & Sons, 1996.

[27] W. J. Conover, Practical Nonparametric Statistics. John Wiley & Sons, 1999.

[28] D. R. Causton, A Biologist's Advanced mathematics. London: Allen & Unwin, 1987.

[29] I. T. Jolliffe, Principal Component Analysis. New York: Springer-Verlag, 1986.