Application of the Fuzzy C-Means Clustering Method on the Analysis of non Pre-processed FTIR Data for Cancer Diagnosis

Xiao Ying Wang, Jon Garibaldi, Turhan Ozen
Department of Computer Science and Information Technology
The University of Nottingham, United Kingdom
{xyw,jmg,txo}@cs.nott.ac.uk

Abstract

Fourier-transform infrared spectroscopy (FTIR) is an efficient, sensitive and computer operated technique that can detect changes in cellular composition that may reflect the onset of a disease. As such, it is being investigated as a method for automatic early detection of pre-cancerous changes. In previous work, FTIR spectral data was first empirically pre-processed and then classified using various data clustering techniques in order to compare to manually obtained classifications. It was found that accurate clustering could only be achieved by manually applying pre-processing techniques that varied according to the particular sample characteristics. In this paper, two data clustering techniques, Hierarchical Cluster Analysis (HCA) and Fuzzy C-Means (FCM) clustering, are used to classify sets of oral cancer cell data without a pre-processing procedure. The performances of these two techniques are compared and their differences are discussed. The FCM method was found to perform significantly better.

1. Introduction

As a major health problem to human, cancer has become a main research area for science researchers in all over the world. There are over 200 different cancer types that have been found so far. In Britain, the lifetime risk of developing cancer is more than one in three. The detection of early invasive cancer is essential in reducing mortality rate. Fourier-transform infrared spectroscopy (FTIR) technology has been recently developed to study biomedical conditions and used as a diagnostic tool for various human cancers and other diseases [1-5]. This technology is based on Fourier-transform infrared spectroscopy. Different functional groups of chemical compounds absorb infrared radiation (IR) at characteristic frequencies and the intensities of IR bands depend on their concentration. This technology can detect changes in cellular composition that reflect the onset of a disease and changes in intermolecular interactions in cells. This makes it a potentially powerful tool in cancer diagnosis, as it can help detecting abnormal cells at molecular levels that occur before the change in morphology seen under the light microscope. An advantage of the FTIR technique is that it may be fully automated and hence be less time-consuming than visual inspection. For one sample, measuring the spectrum on FTIR equipment only takes approximately one minute. In addition, it is a sensitive, computer-operated system. Very small amounts of samples are adequate and they can be studied regardless of the sample’s form and physical state. These attributes make the technique of significant potential interest to large scale screening procedures, such as routine screening of cervical smears [6].

The FTIR technique is based on spectral parameters that reflect the structural changes at the molecular level. If the characteristic spectrum of an abnormal and normal tissue component is known, it may be possible to compare the spectra in each cluster to these reference spectra and hence achieve accurate diagnosis.

In previous work [12], FTIR spectral data was first empirically pre-processed and then classified using various data clustering techniques in order to compare to manually obtained classifications. It was found that accurate clustering could only be achieved by manually applying pre-processing techniques that varied according to the particular sample characteristics. In this paper, two data clustering techniques, Hierarchical Cluster Analysis (HCA) [7] and Fuzzy C-Means (FCM) [8] clustering, are used to classify sets of oral cancer cell data without a pre-processing procedure. The performance of these two techniques is presented and their differences are discussed. The results are presented in comparison with a previous study on the same data where the data was pre-processed empirically before a diagnosis analysis. The aim of this research is to develop an advanced cancer diagnosis system that is easy to use, reliable and efficient.

2. Materials and Methods

The HCA method is a statistical method for finding relatively homogeneous clusters of cases based on measured characteristics. It starts with each case in a separate cluster and then combines the clusters sequentially, reducing the number of clusters at each step until only one cluster is left. The Figure 1 shows an example of a dendogram for HCA classification.

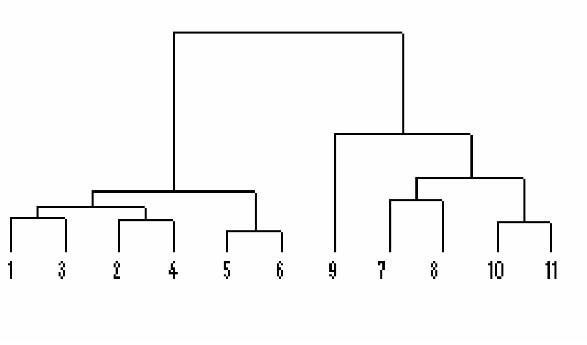

Figure 1: A sample of dendogram for HCA for classifying a set of data points in a plane.

FCM method, also known as Fuzzy ISODATA, which was originally introduced by Bezdek in 1981 as an extension to Dunn's algorithm [9] is the most widely used fuzzy clustering algorithm in practice.

FCM is a data clustering technique based on optimising the objective function:

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

It requires every data point in the data set to belong to a cluster to some membership degree. The purpose of the FCM is to group data points into different specific clusters. Let $X = \{x_1, x_2, ...x_N\}$ be a collection of data. By minimising the objective function (1), X is classified into c homogeneous clusters. Where μ_{ij} is the membership degree of data x_i to a fuzzy cluster set v_j, $V = \{v_1, v_{2,...} v_C\}$ are the cluster centres. $U = (\mu_{ij})_{N*C}$ is a fuzzy partition matrix, in which each μ_{ij} indicates the membership degree for each data point in the data set to the cluster j. The value of U should satisfy the following conditions:

$$\mu_{ij} \in [0,1], \quad \forall i = 1,...N, \forall j = 1...C \quad (2)$$

$$\sum_{j=1}^{C}\mu_{ij} = 1, \quad \forall_i = 1,...N \quad (3)$$

The $\| x_i - v_j \|$ is the Euclidean distance between x_i and v_j. The parameter m is called fuzziness index, which control the fuzziness of membership of each datum. The goal is to iteratively minimise the aggregate distance between each data point in the data set and cluster centres until no further minimisation is possible.

The whole FCM process can be described in the following steps.

Step 1: Initialise the membership matrix *U* with random values, subject to satisfying conditions (2) and (3).
Step 2: Calculate the cluster centre *V* by using following equation

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

Step 3: Get the new distance:

$$d_{ij} = \| x_i - v_j \|, \forall i = 1,..., N, \forall j = 1,...C \quad (5)$$

Step 4: Update the Fuzzy partition matrix *U*:
If $d_{ij} \neq 0$ (means $x_i \neq v_j$)

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C}(\frac{d_{ij}}{d_{ik}})^{\frac{2}{m-1}}} \quad (6)$$

Else $\mu_{ij} = 1$

Step 5:
If the termination criteria have been reached, then stop.
Else go back to step 2.

The suitable termination criteria can be set by checking whether the objective function is below a certain tolerance value or if its improvement compared to the previous iteration is below a certain threshold. Moreover, the maximum number of iteration cycles can be used as a termination criterion as well. [11]

In this study, HCA and FCM clustering techniques were applied to a number of previously obtained clinical data sets with 'gold-standard' classifications by obtained through conventional cytology. Seven sets of FTIR data containing tumour (neoplasm), stroma (connective tissue), 'early keratinisation' and 'necrotic' specimens from three oral cancer patients were provided by Derby City General Hospital to carry out this study. Figure 2 shows an example of FTIR spectra for one of the data sets. A separate analysis of the same data had previously been carried out at Derby City General Hospital. In the report [12] of the previous study, the FTIR data analysis was performed by using Infometrix Pirouette, multivariate analysis software (Infometrix, Inc.). Pre-processing on the spectra was carried out empirically; all spectral range in this study was limited to the 900-1800cm^{-1} interval. The HCA and FCM clustering results obtained during the research presented in this paper are compared with the results presented in the report mentioned above and number of disagreements of classifications is used to compare the performances of HCA and FCM techniques.

In this study, HCA and FCM analysis were performed using MATLAB (version 6.5.0, release 13.0.1).

3. Results

The tumour, stroma, early keratinisation and necrotic specimens were classified using HCA and FCM clustering techniques on seven FTIR data sets. At this stage of our study, we are concerned with asserting spectral characteristic of essentially distinct classes of tissue cells, rather than gradation process or mixed types. So the boundary region points are excluded from the data sets 4 and 7 because the stroma region is invaded by tumour within the vicinity of the boundary between two layers.

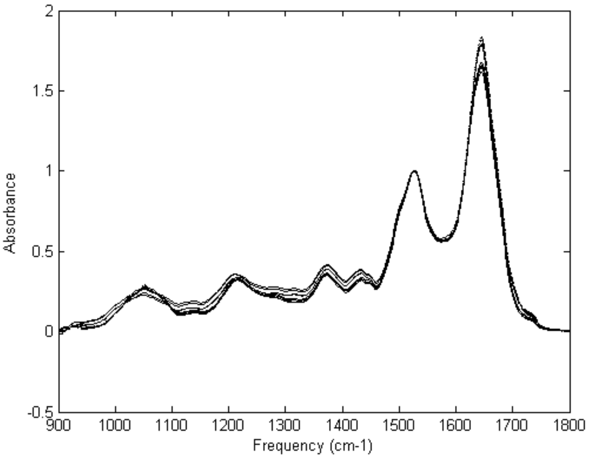

Figure 2: FTIR spectra for data set 1

The comparison results shown are based on the numbers of classifications that are different than those identified in a previous clinical study carried out at Derby City General Hospital. From data set 1 to data set 4, all the data is taken from the first patient, the corresponding data series number is from #001-#092. Data in set 5 and set 6 is taken from the second patient, the corresponding data series number is from #101-#145. Data set 7 is taken from the third patient, the corresponding data series number is from #201-#255. The number of data in each data set (after excluding the boundary data points in data set 4 and 7) is 15, 18, 11, 31, 30, 15 and 42 respectively. Tables 1a -7a show the number of actual data samples of tumour, stroma, early keratinisation and necrotic, and the numbers obtained by using each of the different analysis methods. For instance, in data set 2, 10 data samples were clinically identified as tumour and 8 data samples were identified as stroma. By using the HCA technique, the number of data deemed to be tumour is 17, and 1 is stroma. By using the FCM technique to classify the same data set, the number of data in the cluster deemed to be tumour is 9, in stroma 9, and so on.

Tables 1a - 7a give a general view on the FCM and HCA classification results. In most of data sets, the numbers of data points belonging to tumour, stroma, early keratinisation and necrotic do not exactly match the results from the clinic study. Some HCA classification results even get extreme differences with the clinical results in some data sets (such as in data set 2, 5 and 7). The reason to make the difference is some data points which should be considered in tumour cluster are missed, and at the same time, those missed data points are misclassified into the stroma cluster as extra data points, and vice versa. For example, in data set 2, by using HCA technology, the numbers of data points considered as tumour is 17, while 1 is considered as stroma. Actually, in the stroma cluster, 7 data points are missed, meanwhile these missed data points are misclassified into tumour cluster as extra data points. We will regard these extra data points from HCA and FCM cluster technologies as the number of disagreement of classification to compare with the results from previous clinic study results. Tables 1b - 7b show the comparison results of FCM and HCA classifications based on the disagreements with the results of the previous clinical study at the Derby City General Hospital.

4. Discussion

From the results presented in the previous section, it can clearly be observed that the number of disagreements of classification by the FCM technique is less than that by the HCA technique. In simple terms, the FCM technique appears to achieve better classification than HCA. Table 8 shows the total number of disagreement of classifications by HCA and FCM methods. Over all the data sets, the HCA incorrectly classifies more than twice as many data samples as FCM. Nevertheless the FCM technique can produce a fairly high number of disagreements of classifications in some data sets (i.e. data set 3 and 4). Based on experimental results, we have obtained, we can put forward two hypotheses.

Table 1a: Data set 1 - distribution number of tumour, and stroma in clinic study, HCA and FCM.

Data set 1	Clinic Study	HCA	FCM
Tumour	10	10	10
Stroma	5	5	5

Table 1b: Data set 1 - comparison results based on number of disagreement of classifications by HCA and FCM techniques.

Data set 1	HCA	FCM
Tumour	0	0
Stroma	0	0

Table 2a: Data set 2 - distribution number of tumour, and stroma in clinic study, HCA and FCM.

Data set 2	Clinic Study	HCA	FCM
Tumour	10	17	9
Stroma	8	1	9

Table 2b: Data set 2 - comparison results based on number of disagreement of classifications by HCA and FCM techniques.

Data set 2	HCA	FCM
Tumour	7	0
Stroma	0	1

Table 3a: Data set 3 - distribution number of tumour, and stroma in clinic study, HCA and FCM.

Data set 3	Clinic Study	HCA	FCM
Tumour	8	4	4
Stroma	3	7	7

Table 3b: Data set 3 - comparison results based on number of disagreement of classifications by HCA and FCM techniques.

Data set 3	HCA	FCM
Tumour	0	0
Stroma	4	4

Table 4a: Data set 4 - distribution number of tumour, stroma and early keratinisation in clinic study, HCA and FCM.

Data set 4	Clinic Study	HCA	FCM
Tumour	12	19	11
Stroma	7	5	8
Early Keratinisation	12	7	12

Table 4b: Data set 4 - comparison results based on number of disagreement of classifications by HCA and FCM techniques.

Data set 4	HCA	FCM
Tumour	7	3
Stroma	5	4
Early Keratinisation	0	0

Table 5a: Data set 5 - distribution number of tumour, and stroma in clinic study, HCA and FCM.

Data set 5	Clinic Study	HCA	FCM
Tumour	18	1	14
Stroma	12	29	16

Table 5b: Data set 5 - comparison results based on number of disagreement of classifications by HCA and FCM techniques.

Data set 5	HCA	FCM
Tumour	0	0
Stroma	17	4

Table 6a: Data set 6 - distribution number of tumour, and stroma in clinic study, HCA and FCM.

Data set 6	Clinic Study	HCA	FCM
Tumour	10	10	10
Stroma	5	5	5

Table 7a: Data set 7 - distribution number of tumour, stroma and necrotic in clinic study, HCA and FCM.

Data set 7	Clinic Study	HCA	FCM
Tumour	21	28	18
Stroma	14	13	16
Necrotic	7	1	8

Table 8: Total number of disagreement of classifications by HCA and FCM techniques.

	HCA	FCM
Total	48	19

First in HCA, each observation represents a single cluster initially, and then clusters are merged at each until only one cluster remains. The number of clusters is determined subjectively. If one of the observations is erroneous (e.g. that section cell is damaged), then the whole classification will be affected. For instance, in data set 2, if we eliminate 33rd datum and do HCA analysis again, the number of disagreements of classification in that data set becomes zero. That is, the HCA method is very sensitive to what is probably erroneous data.

Secondly, the FCM technique appears to achieve better classification when the cluster's size and shape are approximately the same. In the given data sets, some clusters sizes have big differences. For example, in data set 3. The number of data belonging to tumour is 8, belonging to stroma is 3; FCM gets 4 disagreements compared with the gold-standard classification. However in data set 2, the number of data belonging to tumour is 10, belonging to stroma is 8; FCM gets 1 disagreement compared with the known classification. Dae-Won et al [10] propose a fuzzy cluster validation index based on inter-cluster proximity for solving this problem.

Table 6b: Data set 6 - comparison results based on number of disagreement of classifications by HCA and FCM techniques.

Data set6	HCA	FCM
Tumour	0	0
Stroma	0	0

Table 7b: Data set 7 - comparison results based on number of disagreement of classifications by HCA and FCM techniques.

Data set7	HCA	FCM
Tumour	7	0
Stroma	0	2
Necrotic	1	1

Overall, these results indicate that FCM is a promising clustering technique for analysing the non pre-processed FTIR data. In work currently underway, we are investigating methods to extend the clustering process into a classification (diagnostic) process. The result of the clustering process is simply to split the data into two or more unlabelled categories. In the results presented above, the clusters were mapped to the actual known classifications in such a way as to minimise mis-classifications in each case. In a true diagnostic process it is necessary to be able to give the clinical user a predicted class for a novel (previously unseen) data sample. The ability to accurately cluster data samples according to the known classifications is the first step in establishing whether the FTIR technique is likely to be clinically useful.

The ultimate goal of this research is to establish the techniques necessary to develop clinically useful tools in a number of clinical domains: e.g. oral cancer, cervical smear test screening, etc. Currently, this ambitious goal remains a long way off. Further research for improving the FCM technique in order to create an advanced cancer diagnosis system is ongoing. We are attempting to obtain significantly larger numbers of samples of known classification from a wider range of patients, and are also in the process of extending the type of samples to sources other than the oral samples presented here. In particular, we are actively engaged on a research programme to obtain FTIR spectral data from cervical smear test samples. Other research challenges include developing the experimental techniques necessary to obtain reliable spectra from the nucleus of a single cell, and the development of the analytical techniques necessary to aggregate the classifications of multiple single nucleus spectra from a given patient into an overall diagnosis.

5. Conclusion

FTIR technology based on detecting abnormal cells at molecular levels has become a prominent technology in cancer diagnosis in recent years. In previous work, the FTIR spectral data has been first empirically pre-processed prior to classification using various data clustering techniques. In order to avoid the extra tools, time and expertise necessary for the pre-processing procedure, it is desirable to avoid pre-processing. In this study, HCA and FCM techniques are used to classify sets of oral cancer cell data without a manual pre-processing procedure. The performance of these two techniques is presented and their results, which are based on the numbers of classifications that are different than those identified in a previous clinical study, are shown. The FCM method was found to perform significantly better and further research on improving this method is on going.

Acknowledgements

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.

References

[1] Wong PTT, Rigas B. Infrared spectra of microtome sections of human colon tissues. *Applied Spectroscopy,* 44:1715–8, 1990.

[2] Rigas B, Morgello S, Goldan IS, Wong PT. Human colorectal cancers display abnormal Fourier-transform infrared spectra. *Proceedings of the National Academia of Science USA,* 87(20):84–8, 1990.

[3] Wong PTT, Goldstein SM, Grekin RC, Godwin TA, Pivik C, Rigas B. Distinct infrared spectroscopic patterns of human basal cell carcinoma of the skin. *Cancer Research*, 53(4):762–5, 1993.

[4] Morris BJ, Lee C, Nightingale BN, Molodysky E, Morris LJ, Appio R. Fourier transform infrared spectroscopy of dysplastic, papillomavirus-positive cervicovaginal lavage speciens. *Gynecological Oncology*, 56(2):245–9, 1995.

[5] Malins DC, Polissar NL, Nishikida K, Holmes EH, Gardner HS, Gunselman SJ. The etiology and prediction of breast cancer. Fourier transform-infrared spectroscopy reveals progressive alterations in breast DNA leading to a cancer-like phenotype in a high proportion of normal women. *Cancer*, 75(2):503–17, 1995.

[6] Lowry SR. The analysis of exfoliated cervical cells by infrared microscopy. *Cellular and Molecular Biology*, 44(1):169-77, 1998.

[7] Gong, X., and M. B. Richman, 1995: On the application of cluster analysis to growing season precipitation in North America east of the Rockies. *Journal Climata*, 8:897-931, 1995.

[8] Bezdek J, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, USA, 1981.

[9] Dunn JC, “A fuzzy relative of the ISODATA process andits use in detecting compact, well separated clusters”, *Journal of Cybernetics*, 3(3):32-57, 1974.

[10] Dae-Won Kim, Kwang H. Lee, Doheon Lee, Fuzzy cluster validation index based on inter-cluster proximity. *Pattern Recognition Letters*, 24: 2561-74, 2003.

[11] Lampinen T, Koivisto H, Honkanen T, Profiling Network Applications with Fuzzy C-Means Clustering and Self-organizing Map. *Proceeding of 1st International Conference on Fuzzy Systems and Knowledge Discovery. Orchid Country Club, Singapore.* 1:300-4, 2002.

[12] Allibone R, Chalmers JM, Chesters MA, Fisher S, Hitchcock A, Pearson M, Rutten FJM, Symonds I, Tobin M, FT-IR microscopy of oral and cervical tissue samples. *Internal Report*, Derby City General Hospital, 2002.