

An Investigation into the use of Fuzzy C-Means Clustering of Fourier Transform Infrared Microscopic Data for the Automation of Breast Cancer Grading

Shabbar Naqvi Jonathan M. Garibaldi
Intelligent Modelling and Analysis (IMA) Research Group
School of Computer Science
University of Nottingham, NG8 1BB, UK
szn@cs.nott.ac.uk *jmg@cs.nott.ac.uk*

Abstract

Medical prognosis plays a vital role in predicting the survival of patients. In the case of breast cancer, prognostic indices are frequently used for estimating survival of patients. One of the most widely used indices is the Nottingham Prognostic Index (NPI) which combines three independent factors for prognosis. Of these factors, grading is an important parameter and is determined subjectively by Nottingham Grading System. There is need for development of automated methods to help pathologists in determining the grade of breast cancer. One such potential technique, Fourier Transform Infrared Micro-Spectroscopy, has been investigated by many researchers in different areas, including the medical sciences. This paper describes the initial work done to apply this technique with advanced computational methods as an objective replacement of the current breast cancer grading system.

1 Introduction

Breast Cancer, which has the highest incidence rate in women, is also the most common cancer in UK. More than one million people are diagnosed with breast cancer every year around the world [1]. After the disease has been diagnosed, monitoring its progress with the passage of time and approximating the re-occurrence of disease based on the complication of the disease for better prediction of survival of patients is very important [2]. This approximation is, in general, known as prognosis which has a significant role in predicting patients' survival. In estimating long term survival, prognostic indices have shown good performance [3]. One of the most widely used indices is the Nottingham Prognostic Index (NPI), which considers tumour diameter,

lymph node status and tumour grade as parameters for prognosis. Of these three factors, grading is one of the most crucial parameters and is determined by the Nottingham Grading System (NGS) which is based on the microscopic evaluation of tumour cells by pathologists [4, 5]. This microscopic evaluation is dependent upon observer variation and variability based on the spatial focus of observation which may result in variable prognosis and sub-optimal treatment [6, 7]. Therefore, there is need to find automated methods that can help pathologists to determine the correct breast cancer grade or perhaps even fully automate the process in a completely objective manner.

The long-term aim of the work described in this paper is to investigate the potential of Fourier Transform Infrared Micro-Spectroscopy (FTIR Microscopy) in providing an automated method for finding the grade of breast cancer while comparing the results with the NGS, with the help of advanced computational methods. The motivation behind this is the fact that FTIR Microscopy has been increasingly applied to the study of biomedical conditions that includes differentiating between different types of cancer tissues [8-15]. To the best of our knowledge, there has been no reported work carried out on automation of breast cancer grading with FTIR Microscopy based on the NGS criteria.

This paper provides a brief introduction to the NPI, the NGS and its parameters, previous work carried out on automation of the NGS, FTIR Microscopy and the various procedures that have been used for its analysis. Due to the fact that the FTIR data is of very high dimensionality, we also discuss the use of Principal Component Analysis (PCA) for the reduction of this dimensionality. An example of an artificial cancer data set will be used with FTIR, PCA and an unsupervised learning clustering algorithm (Fuzzy C-means Clustering) to illustrate the steps that will be used once novel real FTIR breast cancer spectra data sets of various grades have been obtained.

2 Nottingham Prognostic Index (NPI)

In the 1970's and 1980's, a team of clinicians from the Nottingham City Hospital developed a prognostic index for breast cancer based on tumour diameter, lymph node status and tumour grade [16]. It was subsequently validated and called the Nottingham Prognostic Index (NPI) [4]. It is calculated as:

$$\text{NPI} = (0.2 * \text{tumour diameter in cm}) + \text{lymph node stage} + \text{tumour grade}$$

The possible values of lymph node stage are:

- 1: no nodes affected
- 2: up to 3 nodes are affected
- 3: more than three nodes are affected

The possible values of grade are:

- 1: less aggressive appearance of tumour
- 2: intermediate appearance of tumour
- 3: more aggressive appearance of tumour

The NPI categorises patients into three different prognosis groups. The values generally accepted are *Good* (NPI < 3.4), *Intermediate* (3.4 ≤ NPI ≤ 5.4) and *Poor* (NPI > 5.4). The higher the values of NPI, the less are the chances of survival of patients. The NPI has been recognized as the only properly (externally) validated prognostic index for breast cancer [5].

2.1 Nottingham Grading System (NGS)

The Nottingham grading system is widely used around the world for grading the breast cancer tumours and is also a component of the NPI [5, 17]. It is based on *Mitotic Count*, *Tubule Formation* and *Nuclear Pleomorphism*. The values generally accepted can be summarized in Tables 1-3.

Scores of these three factors then are added together to find the overall grade of cancer as shown in Table 4 [18]. This overall grade score is then used within the NPI.

The high grade patients are considered critical and their chances of survival are also poor. The calculation of grade by a pathologist, keeping in mind all the parameters described, is a complex and time consuming process with high risk of human error. Even the best pathologists have been shown to exhibit variability in their scoring of grade. Therefore, there is a need to develop automated methods for this complex problem that minimise the risks of false prediction of grade.

Table 1: The Tubule Formation parameter of NGS

Criteria	Score
Majority of Tumour (>75%)	1
Moderate Degree (10-75%)	2
Little or None (<10%)	3

Table 2: The Mitotic Count parameter of NGS

Criteria	Score
0-9 Mitoses /10 hpf *	1
10-19 Mitoses /10hpf	2
20 or more Mitoses/10 hpf	3

* hpf: high power field

Table 3: The Nuclear Pleomorphism parameter of NGS

Criteria	Score
Small regular uniform cells	1
Moderate Nuclear size and variation	2
Marked Nuclear variation	3

Table 4: Overall breast cancer grade by summation of all scores

Grade	Combined Score
Low Grade (1)	3-5
Intermediate Grade (2)	6-7
High Grade (3)	8-9

3 Previous Work Done on Automation of the NGS

Tutac et. al [19] proposed a method for determining the NGS based on a semantic indexing technique according to a rule based decision system. They further extended their work by adding a multi-resolution approach of image analysis featuring a Gaussian model [20]. Initially, data of six patients was used for the analysis. The initial results showed that grading scores produced by the method were on the lower side and no clear explanation for this low performance was given by the authors. They stressed the need for detailed experimentation to find an optimal solution to the problem. They also reported that it was the first work done for the automation of breast cancer grade using complete NGS criteria. We will explore the use of FTIR Microscopy in finding an automated method for breast cancer grading.

4 FTIR Microscopy

FTIR Microscopy is based on the principle that when an infrared (IR) beam is passed through a sample, the functional groups within the sample absorb the infrared radiation and the rest of the radiation passes through. The resulting spectrum represents the molecular absorption and transmission. Thus, it creates a molecular fingerprint of the sample, as no two unique molecular structures produce the same infrared spectrum [21]. If the characteristic spectrum of a sample under analysis is known (for example, in a 'fingerprint library'), it may be possible to compare each of the obtained spectra to the reference spectra within the fingerprint library and find the correct result. In the case of cancer tissue samples, this technique finds in depth molecular differences within cells that may result in an earlier detection of cell abnormalities. This technique is very sensitive. Therefore, even with small samples, meaningful and reliable results can be achieved. Therefore, there is good chance that this technique may be effective in differentiating between different grades of breast cancer. It is this issue that the current research seeks to explore.

4.1 Analysis of FTIR Microscopic Data

The result of FTIR is an infrared absorbance spectrum (wave number vs. relative absorbance) as shown in figure 1. The generated spectrum consists of the results of observations of many different variables (wave numbers) for a number of individuals (objects). Each variable may be regarded as constituting a different dimension. If there are n -variables (IR bands), each object may be said to reside at a unique position in an abstract entity referred to as an n -dimensional hyperspace. This hyperspace is necessarily difficult to visualize. There is a need for simplification, or dimensionality reduction, for proper data analysis [21].

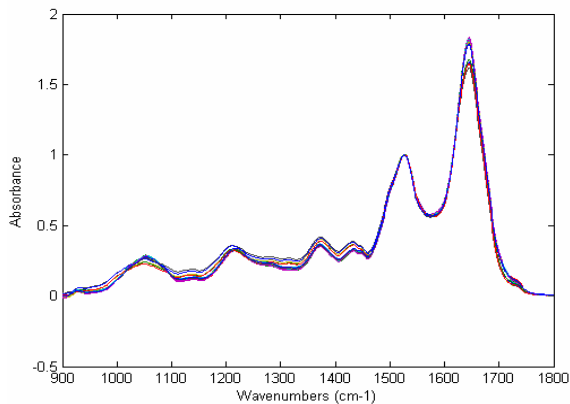


Figure 1[15]: An example of FTIR absorbance spectra

5 Principal Component Analysis

Principal component analysis (PCA) involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. The first principal component (usually called PC1) contains the maximum variation of the data and the subsequent principal component (PC2) contains the next highest amount of variation, etc. In this way, a set of principal components may contain the maximum features of the data sets [22]. As it is an unsupervised data transformation process, no priori knowledge is required before analysis. The computational overhead of the subsequent processing stages is reduced because of the reduction in the number of dimensions. [23].

PCA is widely used for the reduction of high dimension data sets. There are other methods for this purpose such as Independent Component Analysis (ICA), Factor Analysis etc. We have selected PCA for the initial investigations because it has shown good performance on FTIR data sets especially in a previous similar work on breast cancer tissue samples consisting of 7497 x 821 (wave numbers x absorbance)[15]. In future, on newly obtained high dimensional data sets, we will compare the results of PCA with other reputed methods in order to identify the method most suited to FTIR data sets of this kind.

Due to the complex nature of biological systems (like cancer), multivariate clustering algorithms along with PCA have also been frequently applied for the analysis of FTIR spectral data sets [24].

6 Clustering

Clustering is defined as an organization of data points of multidimensional space into groups called clusters based on some similarity criteria. Patterns or points belonging to one cluster are more close to each other than to another cluster. Clustering is an unsupervised learning method as the aim is to group unlabelled patterns into meaningful clusters. Clustering has been used in very many areas like data mining, pattern classification, image analysis etc. [25]. For the current work, we have selected the fuzzy c-means clustering algorithm for the initial investigation. In previous similar work, this algorithm was shown to achieve good results in differentiating between different cancer tissues [13, 14].

6.1 The Fuzzy C-means Clustering Algorithm

The fuzzy c-means (FCM) clustering algorithm developed by Dunn in 1973 and improved by Bezdek in 1981 is commonly used in pattern recognition problems [26, 27]. The aim of the FCM Clustering

algorithm is to minimize the fuzzy objective function as described by the equation

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

where $X = \{x_1, x_2, \dots, x_n\}$ is the collection of data, $V = \{v_1, v_2, \dots, v_c\}$ is the collection of cluster centres, U is the fuzzy partition matrix, μ_{ij} is the membership degree of member x_i to the cluster centre v_j , where $\mu_{ij} \in [0, 1]$ and $\sum \mu_{ij} = 1$, and m is the ‘fuzziness index’ (fuzziness of the clustering) whose value can be chosen from $(1 \leq m < \infty)$. The m parameter is used for controlling fuzziness of the membership of each data point. A small value indicates less fuzziness (when $m = 1$, FCM is the same as the original k-means clustering algorithm from which it was developed), and a large value indicates more fuzziness, but generally 2.0 is the most commonly used value [27].

This algorithm requires the number of clusters to be given as an input. It also randomly initialises the cluster centres and then tries to minimise the fuzzy objective function until a termination criteria is reached. However, it is not guaranteed to find the global minimum of the objective function and so often is restarted multiple times with different starting conditions. A detailed description of this algorithm can be found in [27].

7 Data and Methods

7.1 Data

An artificial FTIR spectral dataset, derived from real datasets, was used for the initial experiments. This dataset consists of 33 different records obtained from three oral cancer patients. This data set was created by the combination of two data sets previously used in a work on oral cancer tissues provided by Derby General Hospital with full consent of patients [15]. Classification made by the histopathologists between tumour and stroma cells was recognized as original results as shown in table 5. The FTIR spectra for 33 different locations of tissue samples were obtained. These locations can be seen in figure 2. The data points 1-15 belong to one data set and 16-33 from the second data set. The dotted white line separates the regions between tumour and stroma. Data points 1-5, 11-15 and 16-25 belong to the tumour cells whereas points 6-10 and 26-33 belong to the stroma cells. The spectral range was limited to $900-1800 \text{ cm}^{-1}$. The resultant FTIR spectra were similar to those shown in figure 1.

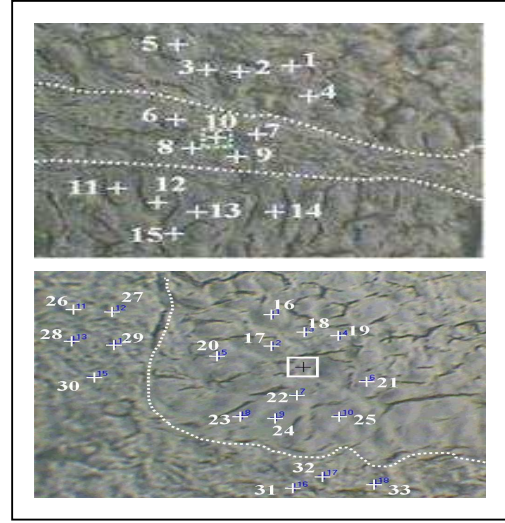


Figure 2 [15]: Location of data points in tissue samples of two data sets used for FTIR analysis

7.2 Methods

A two step approach was used. Firstly, PCA was performed on the data set and then the FCM clustering algorithm was run 10 times on the reduced data set obtained from PCA. For the FCM clustering algorithm, the fuzziness index was set as 2.0 (previous experiments had examined various values of m and found that this was the most effective value [15]), the maximum number of iterations (the termination criteria for the algorithm) was set to 100. For distance measurements, the squared Euclidean distance was used. The number of clusters was set to 2. The variation of the spread of the data was calculated by the method described in [15]. In this method the percentage variance is calculated as

$$100 * \text{sum} [\text{first } n \text{ variances}] / \text{sum} [\text{all variances}]$$

The results were compared with the original results to determine the accuracy. Both PCA and the FCM clustering algorithms were implemented using MATLAB (Version 6.5).

8 Results

The results showed that the above mentioned procedure created two clusters of tumour and stroma having 23 and 10 members respectively. A comparison of results obtained by this experiment with the original results is shown in Table 5.

Table 5: Comparison of results obtained

Dataset	Tissue Types	Original Results	PCA+ FCM
	Tumour	20	23
	Stroma	13	10

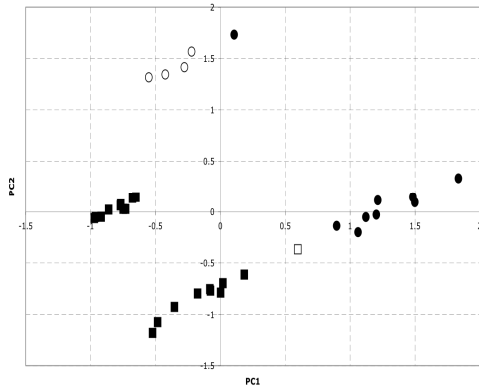


Figure 2: Plot of PC1 and PC2. The squares are actual tumour; the circles actual stroma. The filled (black) symbols are correctly classified; the open symbols incorrectly classified

Table 6: First ten PCs with the associated variance in the data.

Number of PCs	Variance (Percentage)
1	50.77%
2	92.16%
3	97.93%
4	98.65%
5	99.09%
6	99.42%
7	99.57%
8	99.78%
9	99.75%
10	99.81%

These results indicate that there were five data points of the original data that were assigned to the incorrect cluster. In other words, 15.15% of the data was misclassified. This may be due to the adjustments made in the original data sets. The first 10 principal components, along with their percentage variances, are shown in Table 6. This shows that as the number of components increases, the percentage variance also increases (as expected). It also shows that after a certain point this change becomes smaller (for example from PC5 to PC10). The first two PCs contain 92.16% of the variance of the data and their plot gives a clear visual appreciation of the spread of the members of the clusters as shown in figure 2. It shows cluster members of tumour and stroma cells after the execution of PCA and FCM clustering algorithm. It also indicates that more stroma cells (4) were misclassified by the procedure and most of the tumour cells were correctly classified with only one misclassified.

These results indicate that the combination of FTIR spectra data sets with PCA and clustering algorithms may help in better analysis of the data. It

also shows that with a few PCs, maximum information about the data spread can be obtained helping in decreasing computational cost for larger data sets anticipated in the real experiments.

9 Conclusion

This paper has presented the background of the importance of breast cancer grading as a component of the NPI, and has described the complexities involved in the NGS method to determine the grade. The feasibility of using FTIR microscopy to develop an automated method for the objective determination of grade was also explained.

An example of utilising PCA for dimensionality reduction followed by FCM clustering on an artificial FTIR spectra data set was used to show that may indeed be of use in distinguishing between different cancer cells. However, it is clear that there are difficulties in the technique which will need to be overcome if it is to be genuinely useful in clinical practice. This exploratory research is at an early stage, and we remain open-minded as to whether FTIR will be of use in clinical practice.

We are collaborating with the *Breast Cancer Pathology Research Group* at the University of Nottingham for getting new tissue samples and also with the School of Chemistry, University of Nottingham for performing the FTIR analysis of those samples. We are in the process of obtaining new FTIR data sets of breast cancer tumours of various grades to be used in a pilot study for further investigation. These novel data sets will be collected under carefully controlled conditions in order to maximise the chances of determining useful grading algorithms. In future, we intend to investigate both unsupervised learning (different clustering algorithms including hard clustering algorithms) and supervised learning methods (such as fuzzy inference systems) to develop a novel grading method. These methods would be compared to determine the best possible method. The data sets would also be made available once we obtain the real high dimensional data sets for other researchers to work for this highly significant problem. The objective of the current research is to provide a grading method with the help of FTIR microscopy that has high performance, is reliable and is cost-effective, in order to be useful for clinicians in real clinical practice.

References

1. Society, A.C., *Cancer Facts & Figures*. 2008.
2. Ohno-Machado, L., *Modeling Medical Prognosis: Survival Analysis Techniques*. Journal of Biomedical Informatics, 2001. **34**(6): p. 428-39.

3. Soerjomataram, I., W. J. Louwman, M., G. Ribot, J., A. Roukema, J. & Willem W. Coebergh, J., *An overview of prognostic factors for long-term survivors of breast cancer* Breast Cancer Research and Treatment, 2008. **107**(3): p. 309-30.
4. Ellis, I.O., Galea, M., Broughton, N., Locker, A., Blamey, R. W. & Elston, C. W., *Pathological prognostic factors in breast cancer. II. Histological type. Relationship with survival in a large study with long-term follow-up.* Histopathology, 1992. **20**(6): p. 479-89.
5. Rakha, E.A., El-Sayed, M.E., Lee, A.H.S., Elston, C.W., Grainge, M.J., Hodi, Z., Blamey, R.W. & Ellis, I.O., *Prognostic Significance of Nottingham Histologic Grade in Invasive Breast Carcinoma.* J Clin Oncol, 2008. **26**(19): p. 3153-58.
6. Naik, S., Doyle, S., Agner, S., Madabhushi, A., Feldman, M. & Tomaszewski, J., *Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology.* in *Biomedical Imaging: From Nano to Macro*, 2008.
7. Petushi, S., Garcia, F., Haber, M., Katsinis, C. & Tozeren, A., *Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer.* BMC Medical Imaging, 2006. **6**(1): p. 14.
8. Fabian, H., Jackson, M., Murphy, L., Watson, P. H., Fichtner, I. & Mantsch, H. H., *A comparative infrared spectroscopic study of human breast tumors and breast tumor cell xenografts.* Biospectroscopy, 1995. **1**(1): p. 37-45.
9. Fabian, H., Thi, N. A. N., Eiden, M., Lasch, P., Schmitt, J. & Naumann, D., *Diagnosing benign and malignant lesions in breast tissue sections by using IR-microspectroscopy.* Biochimica et Biophysica Acta (BBA) - Biomembranes, 2006. **1758**(7): p. 874-82.
10. Ranjana, M., Deepak Kumar, J., Alka, G. & Kandpal, H. C., *Differentiation of Normal and Malignant Breast Tissues using Infrared Spectroscopy.* 2008: AIP.
11. Rebecca, N.B., Keith, D.R., Neil, S. & Nicholas, S., *Analysis of breast tissue calcifications using FTIR spectroscopy.* 2007: SPIE.
12. Venkatachalam, P., Rao, L. L., Kumar, N. K., Anupama, J. & Shaiju, S. N., *Diagnosis of Breast Cancer Based on FT-IR Spectroscopy.* 2008: AIP.
13. Wang, X.Y., Garibaldi, J. & Ozen, T., *Application of the Fuzzy C-Means Clustering Method on the Analysis of non Preprocessed FTIR Data for Cancer Diagnosis,* in *8th Australian and New Zealand Conference on Intelligent Information Systems.* 2003: Australia. p. 233-38.
14. Wang, X.Y. & Garibaldi, J.M., *Comparison of Fuzzy and Non-Fuzzy Clustering Techniques in Cancer Diagnosis,* in *second international conference in Computational Intelligence in Medicine and Healthcare (The Biopattern Conference).* 2005: Lisbon, Portugal. p. 250-56.
15. Wang, X.Y., *Fuzzy Clustering in the Analysis of Fourier Transform Infrared Spectra for Cancer Diagnosis (PhD Thesis),* in *School of Computer Science.* 2006, University of Nottingham.
16. Haybittle, J.L., Blamey, R.W., and Elston, C.W., *A prognostic index in primary breast cancer.* British Journal of Cancer, 1982. **45**: p. 361-66.
17. Simpson, J.F., Gray, R., Dressler, L.G., Cobau, C.D., Falkson, C.I., Gilchrist, K.W., Pandya, K.J., Page, D.L. & Robert, N.J., *Prognostic Value of Histologic Grade and Proliferative Activity in Axillary Node-Positive Breast Cancer: Results From the Eastern Cooperative Oncology Group Companion Study, EST 4189.* J Clin Oncol, 2000. **18**(10): p. 2059-2069.
18. *Centre for comparative medicine, UC Davis website* Cited 2009; Available from:http://ccm.ucdavis.edu/bcancercd/311/grading_diagram.html.
19. Tutac, A.E., Racoceanu, D., Putti, T., Wei, X., Wee-Kheng, L. & Cretu, V., *Knowledge-Guided Semantic Indexing of Breast Cancer Histopathology Images.* in *International Conference of BioMedical Engineering and Informatics, Sanya, China (BMEI) 2008.*
20. Dalle, J.R., Leow, W. K., Racoceanu, D. & Tutac, A.E., *Automatic Breast Cancer Grading of Histopathological Images.* in *Annual International Conference of the IEEE Engineering, medicine and biology society.* 2008.
21. Ellis, D.I. and Goodacre, R., *Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy.* The Analyst, 2006. **131**: p. 875-85.
22. Jolliffe, I.T., *Principal Component Analysis.* 2 ed. 2002, Aberdeen, U.K.
23. Hyvärinen, A., *Survey on Independent Component Analysis.* Neural Computing Surveys, 1999. **2**: p. 94-128.
24. Wang, X.Y., Garibaldi, J., Bird, B. & George, M., *A novel fuzzy clustering algorithm for the analysis of axillary lymph node tissue sections.* Applied Intelligence, 2007. **27**(3): p. 237-48.
25. Jain, A.K., Murty, M.N., & Flynn, P.J., *Data Clustering: A Review.* ACM Computing Surveys, 1999. **31**(3).
26. Dunn, J.C., *A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters.* Journal of Cybernetics, 1973. **3**(3): p. 32-57.
27. Bezdek, J., *Pattern recognition with fuzzy objective function algorithms.* 1981: Kluwer Academic Publishers Norwell, MA, USA.