

A Case Study in Feature Invention for Breast Cancer Diagnosis Using X-Ray Scatter Images

Shane M. Butler¹, Geoffrey I. Webb¹, and Rob A. Lewis²

¹ School of Computer Science and Software Engineering,
Bldg 26, Monash University Vic. 3800, Australia
{sbutler,webb}@infotech.monash.edu.au

² School of Physics and Materials Engineering,
Bldg 19, Monash University Vic. 3800, Australia
rob.lewis@spme.monash.edu.au

Abstract. X-ray mammography is the current clinical method for screening for breast cancer, and like any technique, has its limitations. Several groups have reported differences in the X-ray scattering patterns of normal and tumour tissue from the breast. This gives rise to the hope that X-ray scatter analysis techniques may lead to a more accurate and cost effective method of diagnosing breast cancer which lends itself to automation. This is a particularly challenging exercise due to the inherent complexity of the information content in X-ray scatter patterns from complex heterogeneous tissue samples. We use a simple naïve Bayes classifier as our classification system. High-level features are extracted from the low-level pixel data. This paper reports some preliminary results in the ongoing development of this classification method that can distinguish between the diffraction patterns of normal and cancerous tissue, with particular emphasis on the invention of features for classification.

Keywords. Knowledge discovery and data mining; Applications.

1 Introduction

The current clinical method used for screening for breast cancer is X-ray mammography. In this method X-rays are directed through the breast and a radiograph is recorded. A radiologist then examines the photograph for evidence of cancer and the appropriate action taken.

Mammography has been successful at reducing breast cancer mortality [7], but like all diagnostic techniques has its limitations. Mammography is non-specific since X-ray attenuation has no direct connection with the presence of disease. As a result it features a high false-positive rate, with less than 20% of women recalled following a suspicious mammogram actually having breast cancer [1]. It also has a significant false-negative rate, with an overall sensitivity (the proportion of cancers successfully detected) of 90% [6]. The sensitivity is further reduced in younger women, in those women with a dense background pattern and in women on hormone replacement therapy [8, 9].

Many women recalled require percutaneous or surgical breast biopsy which is subject to pathological analysis to confirm the nature of their breast lesion. Unfortunately, biopsy analysis is also subject to errors, both from sampling errors, where the sample does not contain part of the lesion, and because the interpretation of a substantial fraction of biopsies is difficult. Whilst the screening programme is undoubtedly successful at detecting cancer, it has important limitations, missing some abnormalities and providing insufficient information for the classification of others. Any technique capable of reducing the need for breast biopsy and/or aiding the analysis of biopsy specimens, especially in the presence of sampling error would be highly advantageous.

Small Angle X-Ray Scattering (SAXS) is a X-ray diffraction based technique where a narrow collimated beam of X-rays are focused on to a sample and the scattered X-rays recorded by a detector. The pattern of the scattered X-rays carries information on the molecular structure of the material. The technique is particularly successful when combined with a synchrotron [12] which produces monochromatic X-ray beams of sufficient intensity to allow scatter patterns to be recorded in a few seconds. The Australian Synchrotron will be operational in 2007.

Small angle X-ray scattering is particularly useful where the material possesses a well ordered molecular structure such as is the case with several biological materials. In particular, collagen which is the most common protein in the human body, is a major constituent of breast tissue and yields an immediately recognisable scatter pattern.

Several groups have applied X-ray scattering to breast tissue [4, 5, 3] and all have detected differences in the scattering patterns produced by normal and tumour tissue. The hope is that this will lead to a more accurate and cost effective method of diagnosing breast cancer.

The challenge remains, however, for a method of automatic classification of the SAXS images. This paper reports on some preliminary results in developing a classification method that can distinguish between diffraction patterns of normal and cancerous tissue, with particular emphasis on the invention of features for classification.

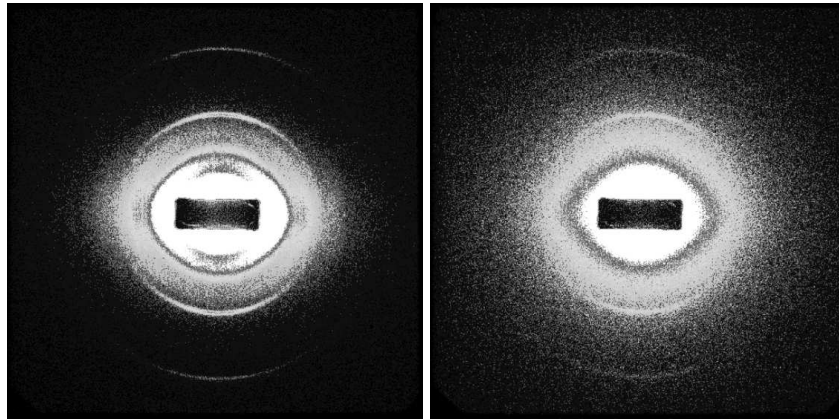
2 Data Description

The data used in this paper has previously been published by Lewis *et al* [5]. Samples were obtained from patients via cosmetic breast reduction and core cuts of invasive breast carcinomas. Diffraction data was then collected at the Synchrotron Radiation Source at Daresbury Laboratory in the UK. All samples were examined and classified by a pathologist. See Lewis *et al* [5] for details of samples and the experimental protocol.

The subset of the data used in this paper consisted of 512 pixels by 512 pixels diffraction images obtained from normal and invasive malignant tumours. Each pixel has a real number associated with it that represents the number of photons

hitting a point on the detector. The classes used were Normal (which consisted of 20 instances) and Tumour (which consisted of 22 instances).

Some example diffraction images are shown in Figure 1. The dark area in the centre is created by the beam stop employed to prevent the direct beam from impinging upon the detector. The scatter pattern radiates from the beam centre which is not necessarily the centre of the beam stop. The well structured collagens found in healthy breast tissue produces clear arcs such as those visible in Figure 1(a). It is believed that many breast tumours express enzymes that degrade collagen [2]. This is consistent with the diffraction patterns from tumour tissue such as that shown in Figure 1(b) where the arcs are much less prominent.



(a) Normal

(b) Tumour

Fig. 1. Diffraction images were obtained using a synchrotron.

3 Feature Extraction

Whilst each pixel could have been treated as an attribute, it is unlikely that a classifier built from such a low-level description would be able to classify the images. Furthermore, there is inherent uncertainty in the scatter of the diffracted X-rays and hence the values at individual pixels cannot be considered as precise. High-level features therefore need to be extracted from the diffraction images. We report here on preliminary work identifying such features.

In order to develop useful features which can subsequently be related to the changes from which the differences arise it is important to develop an understanding of the physics underlying the problem. The angles to which X-rays are

scattered depends on the molecular properties of the material. As a direct result, not only does the pattern change but the amount of X-ray energy hitting the detector varies according to this structure. The first feature seeks to use this possible difference between classes for classification. The value of this feature was sum of all of the pixel intensities across an entire scatter image. We designated it SumIntWhole.

Note that this feature does not make use of the diffraction patterns. It is hypothesised that if the structure of the tissue has been degraded due to cancer, the intensity of the rings or arcs at specific angles will be lower. Locating the rings is relatively straight forward in the case where the structure is intact, but is more problematic when the rings are less well defined. For this reason, and also to search for other significant features, we chose not to directly locate and assess the rings. Rather, we seek to segment the images in such a way as to localise the regions in which the ring may occur. Whether or not one observes a complete ring or an arc such as seen in Figure 1(a) is dependent upon the degree of orientation of the tissue. In the experiments care was taken to align the tissue in the same way from sample to sample but minor variations in orientation will always exist.

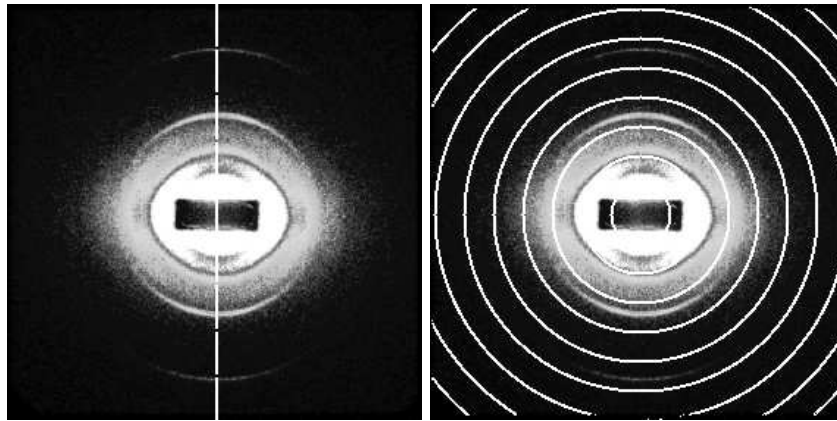
Given the orientation is known to be approximately vertical down the centre of the image, a vertical slice down this line will locate the points at which the highest intensities are expected. There is a further problem however, in that the centre of the beam does not necessarily correspond to the centre of the image or centre of the beam stop. For technical reasons it is difficult to determine where the centre of the beam falls but an attempt has been made to do this, of unknown accuracy. Due to the experimental controls the expectation is that the same location should be the centre of the beam for all samples.

The first method of segmentation used a slice down the diffraction image from $(260, 0)$ to $(260, 512)$ as shown in Figure 2(a). $(260, 259)$ is the estimated beam centre. Since individual intensity values recorded as a pixel by the detector can be spurious, a simple averaging function was employed to ensure that finding the maximum value was less heavily influenced by noise. The calculated value was the average of the pixel and the pixels immediately surrounding it, such that the result was an average of 9 pixel intensities (with the exception of pixels located on the edge of the image that were averages of fewer pixel intensities). The slice was split into either 5 or 10 sections and the maximum (smoothed) value of a pixel in a section_s was recorded in a feature designated MaxSliceSection_s. The y -axis location of the maximum for section_s was also recorded as another feature, YlocMaxSliceSection_s. These features should pick up the differences in intensity between the peaks.

The rings can vary in intensity, distance from the beam centre and how complete they are. To capture this information, a radial segmentation technique was employed. Images were segmented into either 5 or 10 circular regions radiating from the beam centre. This is shown in Figure 2(b). From each region_r, several features were gathered. SumCircularRegion_r and MaxCircularRegion_r are the sum of all intensity values in the region and maximum of the inten-

sity values for these regions, respectively. Intensity values were again averaged for determining the maxima. Following the detection of the maximum intensity for a region $_r$, the distance from that point to the beam centre is designated $RlocMaxCircularRegion_r$.

As the correct beam centre is uncertain, $RlocMaxCircularRegion_r$ may be compromised. The features $YlocMaxCircularRegion_r$ and $XlocMaxCircularRegion_r$ were therefore included as alternate maximum location measures (using x and y co-ordinate values) as they may be less sensitive to this error. Note: Throughout the remainder of this paper feature variable names represent all of the variables relating to that feature, unless otherwise noted.



(a) Centre slice

(b) Circular regions

Fig. 2. Segmentation used among some of the features.

By sampling both the slice sections and circular regions at different frequencies of 5 and 10 we hoped to create features that included the significant information.

4 Classification

A naïve Bayes classifier was selected as the initial base classifier since it is known to be a simple, efficient and effective classifier. These features are all numeric, so we were faced with the issue of whether to use probability density estimation or discretization, and if the latter, which discretization to employ. Due to the small amount of data we deemed probability density estimation inappropriate as accurate estimation of a probability density function may require large volumes

of data. Influenced by Yang and Webb [11] we instead used Equal Frequency Discretization (EFD) with the number of buckets set to 5.

The Waikato Environment for Knowledge Analysis (WEKA) [10] software version 3.3.6 was selected for this analysis as it encompasses both the chosen classifier and discretization method required. This meant that it was necessary for the input data to be written in the WEKA ARFF file format.

The settings used in the experiments were as follows. The `weka.classifiers.bayes.NaiveBayes` classifier was used in conjunction with `weka.filters.unsupervised.attribute.Discretize`. The number of bins was set to 5 and the `useEqualFrequency` option enabled. The leave-one-out cross-validation test method was used and all other settings were the WEKA program defaults.

5 Results

When the naïve Bayes classifier was run on the various features it produced some very interesting results, shown in Table 1. The first column gives the feature a feature number. Most features are the result of splitting some area into either 5 or 10 separate sections. These will have a range in the feature number column instead of just one number, with each value in the range representing a feature corresponding to one section. The second column is the feature name. The accuracy using a naïve Bayes classifier with EFD discretization on only the single feature or group of features is reported in the next column.

Table 1. Leave-one-out naïve Bayes classification accuracy.

Feature #	Name	Accuracy
1	SumInt Whole	45.24%
2-11	MaxSliceSections (10)	76.18%
12-21	YlocMaxSliceSections (10)	78.57%
22-31	SumCircularRegions (10)	90.48%
32-41	MaxCircularRegions (10)	85.71%
42-51	YlocMaxCircularRegions (10)	69.05%
52-61	XlocMaxCircularRegions (10)	80.95%
62-71	RlocMaxCircularRegions (10)	73.81%
72-76	MaxSliceSections (5)	61.90%
77-81	YlocMaxSliceSections (5)	76.19%
82-86	SumCircularRegions (5)	80.95%
87-91	MaxCircularRegions (5)	95.24%
92-96	YlocMaxCircularRegions (5)	42.86%
97-101	XlocMaxCircularRegions (5)	57.14%
102-106	RlocMaxCircularRegions (5)	90.48%

The individual features performed very differently. `SumIntWhole` was the worst performing feature. Similarly, the x and y features at samples of both

5 and 10 sections also delivered very low accuracy. Some of the features that stand out are the 10 SumCircularRegions, the 5 MaxCircularRegions and the 5 RlocMaxCircularRegions, although interestingly for no set of features was equivalent performance obtained using both 5 and 10 partitions.

Table 2. Combinations of the 10 slice sections and 10 circular regions.

Feature #	Excluded Feature	Accuracy
12-71	MaxSliceSections (10)	97.62%
2-11, 22-71	YlocMaxSliceSections (10)	92.86%
2-21, 32-71	SumCircularRegion (10)	95.24%
2-31, 42-71	MaxCircularRegion (10)	95.24%
2-41, 52-71	YlocMaxCircularRegion (10)	92.86%
2-51, 62-71	XlocMaxCircularRegion (10)	95.24%
2-61	RlocMaxCircularRegion (10)	95.24%
2-71	(None excluded)	97.62%

Our expectation was that effective classification would require a combination of features. We first tried the combination of all 10 slice section features and the 10 circular region features. This delivered extremely high accuracy, 97.62%. This means that only one of the examples was misclassified by leave-one-out cross-validation. To evaluate whether all of the 10 slice section and the 10 circular region features were important for this outcome, we next tried all combinations of these features with a single group omitted. Omitting the MaxSliceSections group did not affect the accuracy obtained, but omitting any other single group did, each one resulting in one or two more misclassifications. These results are presented in Table 2.

Table 3. Combinations of the 5 slice sections and 5 circular regions.

Feature #	Excluded Feature	Accuracy
77-106	MaxSliceSections (5)	90.48%
72-76, 82-106	YlocMaxSliceSections (5)	85.71%
72-81, 87-106	SumCircularRegions (5)	92.85%
72-86, 92-106	MaxCircularRegions (5)	88.10%
72-91, 96-106	YlocMaxCircularRegions (5)	90.48%
72-96, 102-106	XlocMaxCircularRegions (5)	90.48%
72-101	RlocMaxCircularRegions (5)	90.48%
72-106	(None excluded)	90.48%

We also tried the combination of all 5 slice section features and the 5 circular region features in a similar fashion. This delivered high accuracy, 90.48% which equates to four misclassified examples. To evaluate whether all of the 5 slice sec-

tion and the 5 circular region features were important for this outcome, we next tried all combinations of these features with a single group omitted. Omitting most of the groups did not affect the accuracy obtained, but when the groups YlocMaxSliceSections and MaxCircularRegions were individually omitted they resulted in several more misclassifications. Omitting the SumCircularRegions group actually resulted in slightly improved accuracy, with one extra correctly classified case. These results are presented in Table 3.

Table 4. Combinations of both the 5 & 10 slice sections and the 5 & 10 circular regions.

Feature #	Excluded Feature	Accuracy
12-106	MaxSliceSection (10)	92.85%
2-11, 22-106	YlocMaxSliceSection (10)	90.48%
2-21, 32-106	SumCircularRegion (10)	95.24%
2-31, 42-106	MaxCircularRegion (10)	95.24%
2-41, 52-106	YlocMaxCircularRegion (10)	92.85%
2-51, 62-106	XlocMaxCircularRegion (10)	90.48%
2-61, 72-106	RlocMaxCircularRegion (10)	92.85%
2-71, 77-106	MaxSliceSection (5)	92.85%
2-76, 82-106	YlocMaxSliceSection (5)	90.48%
2-81, 87-106	SumCircularRegion (5)	92.85%
2-86, 92-106	MaxCircularRegion (5)	95.24%
2-91, 97-106	YlocMaxCircularRegion (5)	95.24%
2-96, 102-106	XlocMaxCircularRegion (5)	95.24%
2-101	RlocMaxCircularRegion (5)	95.24%
2-106	(None excluded)	95.24%

We also tried the combinations of all features at a sample size of both 5 and 10 sections. Again, we found that omitting some groups of features yielded high accuracy results, with all results being over 90%. These results are shown in Table 4.

6 Conclusions and Future Research

Herein we present preliminary results from an exploratory study that seeks to identify features for diagnosing breast cancer from synchrotron X-ray scatter data. This is a particularly challenging exercise due to the inherent complexity of the information content in X-ray scatter patterns from complex heterogeneous tissue samples. Nonetheless, we are encouraged by the high accuracy achieved by our initial simple features.

Some caution is required in interpreting these results. With many different experiments performed, it should be expected that some observed accuracies are unrealistically high due to chance variation. That we obtained above 90% accuracy on all combinations of features relating to 10 slice sections and 10

circular regions does however provide evidence that we are capturing significant regularities in the data.

Our next challenge is to refine the features. A first step will be to more accurately locate the centre of the beam for each image and then more precisely locate and measure each of the rings.

We are also seeking to increase the size of the data set by collecting further samples. This will allow better evaluation of the features. The ultimate test will be to evaluate a specific classifier on a new data set of previously unsighted cases. However, the research is still a long distance short of this objective.

Our preliminary results provide some hope to the prospect of developing X-ray based techniques for detecting and diagnosing cancer that are more accurate and less intrusive than those currently existing. If such techniques could be developed they may lead to further decreases in mortality coupled with a decrease in the intrusiveness and uncertainty to which large numbers of women are currently subjected.

7 Acknowledgments

We would like to acknowledge Marcus Kitchen for his invaluable consultation in interpreting the SAXS data.

We also wish to thank the authors of the WEKA machine learning environment [10], which was used in these experiments, for making their software freely available.

References

1. N. Bjurstam, L. Bjorneld, S. W. Duffy, T. C. Smith, E. Cahlin, O. Eriksson, L. O. Hafstrom, H. Lingaas, J. Mattsson, S. Persson, C. M. Rudenstam, and J. Save-Soderbergh. The gothenburg breast screening trial. *Cancer*, 80(11):2091–2099, 1997.
2. L. M. Coussens and Z. Werb. *Chemistry and Biology*, 3:895–904, 1996.
3. M. Fernandez, J. Keyrilainen, R. Serimaa, M. Torkkeli, M-L. Karjalainen-Lindsberg, M. Tenhunen, W. Thomlinson, V. Urban, and P. Suortti. Small-angle x-ray scattering studies of human breast tissue samples. *Physics in Medicine and Biology*, 47:577–592, 2002.
4. G. Kidane, R. D. Speller, G. J. Royle, and A. M. Hanby. X-ray scatter signatures for normal and neoplastic breast tissues. *Physics in Medicine and Biology*, 44:1791–1802, 1999.
5. R. A. Lewis, K. D. Rogers, C. J. Hall, E. Towns-Andrews, S. Slawson, A. Evans, S. E. Pinder, I. O. Ellis, C. R. M. Boggis, A. P. Hufton, and D. R. Dance. Breast cancer diagnosis using scattered x-rays. *Journal of Synchrotron Radiation*, 7:348–352, 2000.
6. A. I. Mushlin, R. W. Kouides, and D. E. Shapiro. Estimating the accuracy of screening mammography: a meta-analysis. *Am J Prev Med*, 14(2):143–53, 1998.
7. L. Nystrom, L. Rutqvist, and S. Wall et al. Breast cancer screening with mammography: overview of swedish randomised trials. *Lancet*, 341:973–978, 1993.

8. R. D. Rosenberg, W. C. Hunt, M. R. Williamson, F. D. Gilliland, P. W. Wiest, C. A. Kelsey, C. R. Key, and M. N. Linver. The effect of age, density, ethnicity, and estrogen replacement therapy on screening mammography sensitivity and cancer state: A review of 183,134 screening mammograms in Albuquerque, New Mexico. *Radiology*, 209:511–518, 1998.
9. D. M. Sibbering, H. C. Burrell, A. J. Evans, L. J. Yeoman, R. M. Wilson, and J. F. Robertson. Mammographic sensitivity in women under 50 years presenting symptomatically with breast cancer. *Breast*, 4(2):127–129, 1995.
10. I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning with Java Implementations*. Morgan Kaufmann, Sydney, 2000.
11. Y. Yang and G. I. Webb. Discretization for naive-bayes learning: managing discretization bias and variance. Technical Report 2003/131, School of Computer Science and Software Engineering, Monash University, 2003.