MULTIPLE INSTANCE LEARNING FOR BREAST MRI BASED ON GENERIC SPATIO-TEMPORAL FEATURES

Fahira Afzal Maken, Andrew P. Bradley

The University of Queensland
School of Information Technology and Electrical Engineering
St Lucia, QLD 4072, Australia

ABSTRACT

In this paper we investigate multiple instance learning (MIL), using generic tile-based spatio-temporal features, for the classification of benign and malignant lesions in breast cancer magnetic resonance imaging (MRI). In particular, we compare the performance of citation-kNN (CkNN) and conventional kNN against a traditional approach based on bespoke features extracted from a segmented region-of-interest (ROI). Results demonstrate that tile-based CkNN has equivalent performance to ROI-based classification. However, the tile-based approach does not require any domain specific features typically used in breast MRI. This not only has the potential to make tile-based classification robust to inaccuracies in the delineation of suspicious lesions, but also makes it suitable for the detection of suspicious lesions prior to segmentation.

Index Terms— Multiple Instance Learning, Breast MRI, Feature Extraction, Feature Selection.

1. INTRODUCTION

In this paper we evaluate the performance of multiple instance learning (MIL) [17] as a ‘pure’ machine learning approach for the classification of breast cancer. Specifically, we use T2-weighted magnetic resonance imaging (MRI) and dynamic contrast enhanced MRI (DCE-MRI). Here we utilize parametric models and a discrete cosine transform (DCT) [13] as feature extraction techniques. In particular, we compare generic tile-based features against region-of-interest (ROI) based features. We also compare the results with those of [9] where we used MIL for the classification of non-contrast enhanced MRI based only on generic tile-based spatial features.

The traditional approach to discriminate benign and malignant breast lesions involves ROI-based methods. Here, the dataset consists of bespoke (i.e. domain specific) features, extracted from each detected and then segmented lesion. In this way, each lesion becomes a labelled instance in the dataset. The feature vector extracted from each lesion is then individually labelled as either benign (negative) or malignant (positive). The features used in traditional single instance learning (SIL) approaches are based on the intensity, texture and morphology of the segmented lesion [15]. However, lesion margins and shape are strongly dependent on an accurate segmentation, which is a challenging task due to poor signal-to-noise-ratio and faint edges due to partial volume effects. Therefore, lesion delineation is affected by variation or uncertainties in the (semi)-automated lesion segmentation process. Clearly, these variations have the potential to lead to variations in diagnostic outcome.

Multiple instance learning is a relatively new paradigm in supervised learning, which appears to be suitable for many computer aided diagnosis (CAD) related problems, particularly when there is uncertainty regarding the class label given to individual instances. MIL is a semi-supervised approach where each labelled sample is represented as a set (or ‘bag’) of instances. The objective of MIL is then to classify the bag of instances rather than the individual instances. In the context of MIL in image analysis, a bag is a sub-image consisting of multiple instances, where those instances are either individual pixels, square tiles (tile-based MIL) or arbitrary regions of interest (ROI-based MIL) [10]. According to the standard (asymmetric) MIL assumption, a bag is labeled positive if at least one instance in the bag is positive, otherwise the bag is negative [17]. In the tile-based approach, the features are generic in nature rather than specific to breast cancer. Since these features are extracted from small tiles, not segmented ROIs, classification performance is not affected by the accuracy of the segmented regions. This makes tile-based MIL suitable for both diagnostic applications, which classify already detected lesions, and screening applications, which initially detect suspicious lesions.

The purpose of this paper is not to solve breast cancer MRI classification problem, but to evaluate the efficacy of MIL as a ‘pure’ machine learning approach for the diagnosis of breast cancer. When we say ‘pure’ we mean that we use MIL as a generic approach without knowing much about physiology and domain specific features typically used in breast cancer MRI. Rather, we use MIL for classification of breast MRI in the same way as it is applied to solve an arbitrary image classification problem. In other words, we

978-1-4673-6997-8/15/$31.00 ©2015 IEEE

utilize generic features based on their level of discrimination as opposed to (bespoke) application specific features selected on the basis of prior knowledge.

2. EXPERIMENTAL METHODOLOGY

2.1. Dataset

Here we use 'dataset A' previously used in our work on the diagnosis of breast cancer MRI in [9]. This dataset consists of 53 malignant and 24 benign mass-like lesions. In [9] we considered a block of 60×60 pixels from T2-weighted (T2-w) MRI overlapping with a manually segmented ROI as a bag and features extracted from $n \times n$ tiles as instances. In this paper we extend this work to both T2-w MRI and DCE-MRI. The data were labelled and normalized as per [9].

The fat suppressed T1-weighted DCE-MRI images were acquired as five stacks using a 3-dimensional (3D) fast spoiled gradient-echo (FSPGR) sequence (Echo time = 3.4ms, Repetition Time = 6.5ms and flip angle of 10°). The first stack corresponds to baseline pre-contrast images and the remaining stacks comprise of post-contrast images. Each stack was acquired in around 90 seconds with a 45 second delay between the pre-contrast and the first post-contrast stack. The second last stack was acquired in a sagittal orientation. All of the remaining stacks were acquired axially with a field of view of 32cm, a 360 × 360 acquisition matrix, and slice thickness of 1mm. The number of slices ranged from 116 to 182 with a median of 150. Here we use only four stacks which were acquired axially.

2.2. Algorithms

We evaluate the performance of a MIL based k-nearest neighbour (kNN) algorithm called citation-kNN (CkNN) [19]. We have chosen citation-kNN because it involves the optimization of only two parameters (reference neighbours '*k*' and citer's rank '*c*'). Moreover, citation-kNN has been used for solving various MIL problems with high accuracy [10]. To compare the performance of citation-kNN in a SIL paradigm we also select kNN. kNN is a simple, effective and non-parametric technique which has been used extensively [8]. In addition, we compare the performance of CkNN with the results of a more conventional approach described in [6] and [7], because this study was performed on the same dataset. This study investigates the discriminatory power of state-of-the-art ROI-based features from multi-modal MRI using a Random Forest (RF) classifier. In particular, we compare the results with the relevant results of study 2 based on DCE-MRI alone and DCE-MRI combined with T2-w MRI.

We use *Multiple Instance Learning Toolbox*[1] [5], an add-on to PRTools toolbox written in Matlab ®, and *PRTools toolbox* 4.2.0[2].

2.3. Features

In [9] we proposed a generic tile-based MIL approach for the identification and classification of non-contrast enhanced breast cancer MRI. Here we extend this approach to DCE-MRI and include both spatial and temporal features. To extract spatio-temporal features, we decompose 64×64 blocks of images into independent 8×8 tiles for T2-w MRI and 8×8×4 cubes for DCE-MRI. These tiles/cubes represent the instances in each 64×64 or 64×64×4 bag. To obtain this spatio-temporal information we evaluate two approaches: 1) using a 3D-DCT on DCE-MRI alone, which gives spatial information combined with temporal i.e. spatio-temporal. 2) using a 2D-DCT on T2-w MRI to obtain spatial features plus three parameters each from a linear slope model [16] and an empirical model of contrast enhancement [18] to obtain temporal features from DCE-MRI. To reduce computational complexity, we do not fit enhancement models voxel-wise, but rather to the relative enhancement based on the mean of each 8×8 tile from DCE-MRI stack. The equation for relative enhancement for DCE-MRI stack is given in [6].

We have chosen these above mentioned parametric models because they are generic in nature. These models are not derived from pharmacokinetics, i.e. these models do not make assumptions about the relation of concentration of contrast agent and intensity (two compartment flow). Also, the parameters of these models are independent of the density and nature of tissue type [18]. The initial parameter estimates for linear slope model and empirical model of contrast enhancement can be found in [3] and [11] respectively.

While using the DCT as a feature extraction technique, we apply a 3D-DCT to each 8×8×4 cube of DCE-MRI and a 2D-DCT to each 8×8 tile of T2-w MRI. We then perform a 3D-zigzag traversal [4] to select 25 (10%) coefficients. By extracting coefficients in a zigzag manner, the correlation between coefficients is minimized [4] and we extract coefficients in increasing order of frequency [14]. For spatio-temporal features from DCE-MRI and T2-w MRI, we extract 25 features by combining the six temporal features extracted from DCE-MRI with the first 19 2D-DCT coefficients extracted in a zigzag transversal.

For fair comparison with the results of [9], we select 5 features only using the *plus-l-take-away-r* algorithm [12]. We select features in both MIL and SIL based distance metrics. Specifically we use minimum Hausdorff distance for MIL based feature selection and minimum Mahalanobis distance for SIL based feature selection. We select features

[1] http://prlab.tudelft.nl/david-tax/mil.html
[2] http://prtools.org/software

on the training data and estimate performance on independent test data via a 10-fold cross validation (CV) [2] scheme. The performance measure utilized is mean area under receiver operating characteristics curve (AUC) because it estimates the probability of correct ranking [1]. Due to limited amount of data, we optimize the parameters of learners used in 3D-DCT by randomly selecting coefficients from the 2D-DCT on T2-w MRI. This results in an unbiased estimate of parameter values because parameters are tuned solely on T2-w MRI which is entirely independent from T1-w DCE-MRI. However, this will bias the estimated performance of the 2D-DCT plus temporal features based classification.

3. RESULTS

Table 1 shows a comparison of classification performance of CkNN, kNN based on 3D-DCT spatio-temporal features and traditional ROI-based classification with RF using just DCE-MRI [6].

Table 1: Performance of 3D-DCT tile-based features against traditional ROI-based features.

Technique	Learner	Mean AUC
Tile-based MIL	CkNN	0.816 ± 0.047
Tile-based SIL	kNN	0.838 ± 0.010
ROI-based SIL [6]	RF	0.824 ± 0.046

Fig. 1 presents the performance of CkNN and kNN with MIL based feature selection in comparison to SIL based feature selection for 3D-DCT features. It can be seen that MIL based feature selection is important for MIL based classification (with CkNN). Similarly, SIL based feature selection is important for SIL based classification.

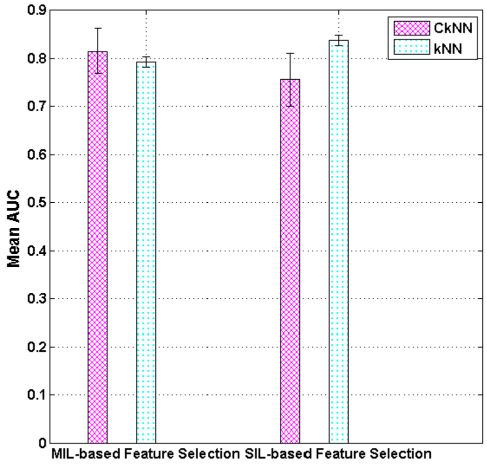

Fig. 1: Performance of learners with MIL based feature selection in comparison to SIL based feature selection for 3D-DCT features.

Table 2 compares the classification performance of CkNN, kNN based on 2D-DCT plus temporal features against traditional ROI-based classification using both DCE-MRI and T2-w MRI with RF [6].

Table 2: Performance of 2D-DCT plus temporal features against traditional ROI-based features.

Technique	Learner	Mean AUC
Tile-based MIL	CkNN	0.778 ± 0.052
Tile-based SIL	kNN	0.702 ± 0.012
ROI-based SIL [6]	RF	0.838 ± 0.045

Fig. 2 shows the classification performance of CkNN and kNN with MIL based feature selection in comparison to SIL based feature selection for 2D-DCT plus temporal features. Fig.2 shows that the MIL based learner (CkNN) performs best with the MIL based feature selection. However, here SIL based classification (with kNN) also gives better performance with MIL based feature selection as compared to SIL based feature selection.

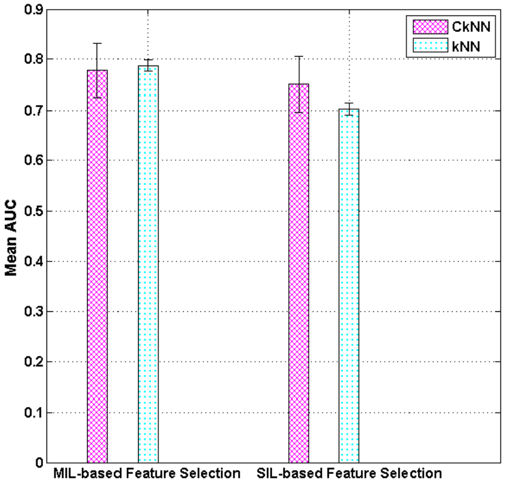

Fig. 2: MIL based feature selection in comparison to SIL based feature selection for 2D-DCT plus temporal features.

4. DISCUSSION

In Table 1 and 2, we have presented the classification of benign and malignant lesions using generic tile-based spatio-temporal features and ROI-based features. A t-test indicates that the performance of CkNN is equivalent to the performance of kNN. Also, generic tile-based classification using CkNN is equivalent to traditional ROI-based classification. Thus, tile based classification using MIL is a viable option for the classification of benign and malignant breast lesions with additional advantages mentioned in [9]. This indicates that generic tile-based features allow us to use MIL as a 'pure' machine learning method to solve the breast MRI classification problem with equivalent performance to the traditional ROI-based classification.

Based on these results, we can say that MIL is a suitable choice for CAD systems. However, a 'pure' machine learning approach has the disadvantage that it produces an unintelligible 'black-box' model to the clinicians. That is, the learning process and generic features are not easily interpretable by the clinicians. However, while traditional ROI-based features may be interpretable by clinicians (as they relate to physiological properties of the lesion), the learning process (i.e. RF) still results in a black-box model.

Fig.3 shows a comparison of the spatio-temporal features from Tables 1 and 2 against the spatial features from [9]. Fig.3 confirms that classification performance is improved with tile-based spatio-temporal features compared to spatial features alone.

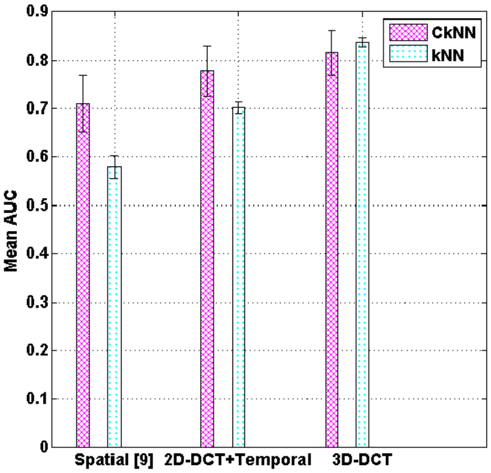

Fig. 3: Performance of spatio-temporal features in comparison to spatial features [9].

Next, we analyse the classification performance of the learners from Table 1 in comparison to Table 2. A t-test indicates that the performance of CkNN with 3D-DCT spatio-temporal features is equivalent to that with 2D-DCT plus temporal features. However, kNN performs significantly better with 3D-DCT features as compared to 2D-DCT plus temporal features. We know that, Table 1 presents an unbised estimate of the performance with 3D-DCT features, while the performance of CkNN and kNN is biased in Table 2 due to parameter optimization on the same dataset. Therefore, the 3D-DCT spatio-temporal features would appear better than 2D-DCT plus temporal features. Moreover with 2D-DCT plus temporal features, we extract temporal information using parametric enhancement models. Although these models are generic, they are domain specific to some extent. Thus we can say that 3D-DCT spatio-temporal features are more 'pure' (generic) than 2D-DCT plus temporal features.

From Fig.1 and 2 we can see that CkNN performs better with MIL based feature selection as compared to SIL based feature selection. This demonstrates importance of MIL based criterion for dimensionality reduction in MIL based classification. Similarly, the SIL based criterion is important for SIL based classification.

After doing feature selection, we assess the relative importance of low and high frequency coefficients selected from the DCT. We first arrange DCT coefficients into frequency groups based on the similar sum of their indices in the same way as is done in [14]. In this way we get five groups for 25 3D-DCT coefficients and six frequency groups for 19 2D-DCT coefficients. Next, we divide DCT frequency groups equally into low and high frequency. We also evaluate the relative importance of horizontal, vertical and diagonal coefficients by counting the number of occurrence of each selected coefficient with respect to its position. In MIL based feature selection, all selected features belong to low frequency group. Moreover, we get equivalent counts of horizontal, vertical and diagonal coefficients. A similar trend was identified for the SIL based feature selection. This demonstrates an approximately equal importance of horizontal, vertical and diagonal low frequency DCT features for classification of mass-like lesions. This statement is in accordance with [9], where we analyzed the relative importance of horizontal and vertical generic tile-based spatial features for the mass-like lesions. For the 2D-DCT plus temporal features, the MIL based feature selection returns only the DC coefficient from the 2D-DCT and four temporal model features. While, with SIL based feature selection, we get all low frequency DCT features and no model features. The high occurrence of model features in MIL based feature selection confirms the importance of temporal information from DCE-MRI for classification of benign and malignant lesions.

5. CONCLUSIONS

In this paper we have evaluated the efficacy of MIL as a 'pure' machine learning technique for the descrimination of benign and malignant lesions in breast MRI. Experimental results indicate that performance of CkNN based on tile-based spatio-temporal features is statistically equivalent to the ROI-based classification. However, the tile-based approach does not require any domain specific features and is robust to inaccuracies in the segmentation of suspicious lesions. Therefore, CkNN may be a suitable choice for the classification of benign and malignant lesions. Also, spatio-temporal features have improved descrimination compared to spatial features alone. Moreover, 3D-DCT spatio-temporal features are better than 2D-DCT plus temporal features. Further, we highlight that MIL based feature selection is important for MIL based classification.

ACKNOWLEDGEMENTS

The approval of the Human Research Ethics Committee at the University of Queensland was obtained for this study. A.P.B. is supported by an Australian Research Council Future Fellowship (FT110100623).

6. REFERENCES

[1] A. P. Bradley, “The use of area under the ROC curve in the evaluation of machine learning algorithms,” *Pattern Recognition*, vol. 30, no. 7, pp.1145 – 1159, 1997.

[2] Ethem Alpaydin, *Introduction to machine learning*, The MIT Press Cambridge: Massachusetts London, England, 2010.

[3] Andrew Mehnert, Kerry McMahon, Dominic Kennedy, Ewert Bengtsson, Stephen Wilson and Stuart Crozier, “Visualization of the pattern of contrast enhancement in dynamic breast MRI,” APRS Workshop on Digital Image Computing (WDIC2005), Brisbane, Australia, 2005.

[4] Boon-Lock Yeo, Bede Liu, “Volume rendering of DCT-based compressed 3D scalar data,” *IEEE Trans. Visualization and Computer Graphics*, vol. 1, no. 1, pp. 29 – 43, Mar 1995.

[5] D. M. J. Tax, MIL: A Matlab toolbox for multiple instance learning, Version 0.8.1, March 2013.

[6] Darryl McClymont, Andrew Mehnert, Adnan Trakic, Dominic Kennedy, Stuart Crozier, “Multimodal features for improved breast MRI CAD,” *Journal of Medical Imaging*, (under review 2).

[7] Darryl G. McClymont, Computer assisted detection and characterization of breast cancer in MRI, Chapter 7, PhD Thesis, 2014.

[8] Dhurandhar A., Dobra A., “Probabilistic characterization of nearest neighbour classifier,” *Int J Mach Learn and Cyber*, vol.4, no. 4, pp. 259 – 272, 2013.

[9] Fahira A. Maken, Yaniv Gal, Darryl McClymont, Andrew P. Bradley, “Multiple instance learning for breast cancer magnetic resonance imaging,” in Proceedings Digital Image Computing: Techniques and Applications (DICTA) , Wollongong, pp. 1 – 8, 2014.

[10] James Foulds and Eibe Frank, “A review of multi-instance learning assumptions,” *Knowledge Engineering Review*, Cambridge University Press 25, no. 01, pp. 1–25, March 2010.

[11] Gal Y., Mehnert A. J., Bradley A. P., McMahon K., Kennedy D., Crozier S., “Denoising of dynamic contrast-enhanced MR images using dynamic nonlocal means,” *IEEE Transactions on Medical Imaging*, vol. 29, no. 2, pp. 302 – 310, 2010.

[12] J. Kittler, “Feature set search algorithms,” *Pattern recognition and signal processing*, The Netherlands: Sijthoff and Noordhoff, pp. 41 – 60, 1978.

[13] K. Rao, P. Yip, *Discrete Cosine Transform: Algorithms, Advantages, Applications*, Academic Press Professional, Inc. San Diego, CA, USA ©1990.

[14] Malavika Bhaskaranand and Jerry D. Gibson, “Distributions of 3D DCT coefficients for video,” International Conference on Acoustics, Speech and Signal Processing (ICASSP), Taipei, pp. 793 – 796, 2009.

[15] N. Bhooshan, M. Giger, L. Lan, H. Li, A. Marquez, A. Shimauchi, and G. M. Newstead, “Combined use of T2-weighted MRI and T1-weighted dynamic contrast-enhanced MRI in the automated analysis of breast lesions,” *Magnetic Resonance in Medicine*, vol. 66, pp. 555 – 64, 2011.

[16] O. Schabenberger and F. J. Pierce, *Contemporary statistical models for the plant and soil sciences*, CRC Press, 2002.

[17] Thomas G. Dietterich, Richard H. Lathrop and Tomas Lozano-Perez, “Solving multiple instance problem with axis-parallel rectangles,” *Artificial Intelligence*, vol.89, no. 1-2, pp. 31 – 71, 1997.

[18] Yaniv Gal, Andrew Mehnert, Andrew Bradley, Kerry McMahon, and Stuart Crozier, “An evaluation of four parametric models of contrast enhancement for dynamic magnetic resonance imaging of the breast,” 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Lyon, France, 2007.

[19] J. Wang and J. D. Zucker, “Solving the multiple-instance problem: a lazy learning approach,” 17th International Conference on Machine Learning, San Francisco, pp. 1119 – 1125, 2000.