

Classification Performance related to Intrinsic Dimensionality in Mammographic Image Analysis

Harry Strange^a and Reyer Zwiggelaar^{a*}

^aDepartment of Computer Science, Aberystwyth University, SY23 3DB, UK

Abstract. In the problem of mammographic image classification one seeks to classify an image, based on certain aspects or features, into a risk assessment class. The use of breast tissue density features provide a good way of classifying mammographic images into BI-RADS risk assessment classes [1]. However, this approach leads to a high-dimensional problem as many features are extracted from each image. These features may be an over representation of the data and it would be expected that the intrinsic dimensionality would be much lower. We aim to find how running a simple classifier in a reduced dimensional space, in particular the apparent intrinsic dimension, affects classification performance. We perform classification of the data using a simple k -nearest neighbor classifier with data pre-processed using two dimensionality reduction techniques, one linear and one non-linear. The optimum result occurs when using dimensionality reduction in the estimated intrinsic dimensionality. This not only shows that optimum performance occurs when classifying in the intrinsic-dimensional space but also that dimensionality reduction can improve the performance of a simple classifier.

1 Introduction

Mammography remains the main tool used for the screening and detection of breast abnormalities and the development of full field digital mammographic imaging systems has led to increased interest in computer aided detection systems [2]. Radiologists are increasingly turning to such Computer Aided Diagnostics (CAD) systems to assist them in the detection and/or evaluation of mammographic abnormalities [3]. As such the reliability and accuracy of such systems is paramount especially as breast cancer constitutes the most common cancer among women in the European Union [4]. Many CAD systems will attempt to detect and classify mammographic abnormalities such as microcalcifications and masses. However there is a strong correlation between breast cancer risk and breast density [5, 6]. Figure 1 shows 4 mammograms covering a range of breast tissue density [1]. Each of these 4 images represents a different BI-RADS class. The American College of Radiology BI-RADS [7] is a widely used risk assessment model. It aims to classify a mammogram into one of four classes according to breast density. The classes can be explained as follows. BI-RADS I: an almost entirely fatty breast, not dense; BI-RADS II some fibroglandular tissue is present; BI-RADS III the breast is heterogeneously dense; BI-RADS IV: the breast is extremely dense. Although BI-RADS is becoming a radiological standard other risk assessment models exist that aim to classify breasts according to different aspects or features present in the mammogram (e.g. Tabár modelling [8]).

For a CAD system to place a mammogram into one of the BI-RADS classes it will need to use some form of classification algorithm. Many algorithms exist for the purpose of classification and generally they work by building a model of the data from “known” examples (i.e. mammograms with known BI-RADS classes). Using this model the classifier will then be able to assign each new mammogram to a BI-RADS class. It is unrealistic to use the raw mammographic

*Email: hgs08@aber.ac.uk, rrz@aber.ac.uk

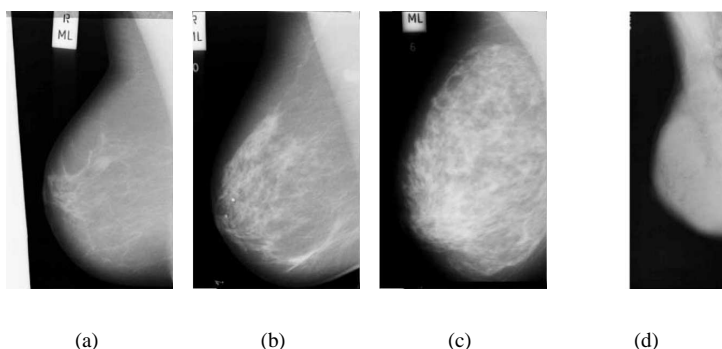


Figure 1. Mammograms showing 4 different breast densities ranging from low density (a) to high density (d).

image as input into a classifier, so features are usually extracted from each image and used as input. In the data from this paper 280 features are extracted from each mammogram (see Section 3). The obvious question to then ask is whether all the extracted features are necessary? Most high-dimensional data will contain redundancy (i.e. dimensions that provide no extra information) which could impair classification performance. Dimensionality reduction is a pre-processing technique that aims to reduce the dimensionality of the data so as to improve classification performance. The question now becomes how many dimensions are needed to best represent the original data? Intrinsic dimensionality estimators aim to find the number of dimensions needed to represent the data without losing important features. In this paper we combine these two elements. We estimate the intrinsic dimensionality of the data and then use dimensionality reduction to show that optimal classification performance occurs when classifying in this intrinsic dimensionality.

We begin by outlining dimensionality reduction and intrinsic dimensionality estimators in Section 2. The data used in this experiment is then discussed in Section 3 before the methodology is outlined in 4. The results are shown in Section 5 before final conclusions and future work are discussed in Section 6.

2 Dimensionality Reduction

Dimensionality reduction is the process of finding from a set of high-dimensional observations a representation of lower dimensionality. This representation will maintain certain aspects, or features, of the original data. Different dimensionality reduction algorithms will retain different features, and this leads to a multi-level taxonomy of techniques. At the highest level techniques can be classified by whether they aim to find a linear subspace within the high-dimensional data, or whether they aim to find a non-linear manifold. We use two dimensionality reduction techniques, one linear (Principal Components Analysis) and one non-linear (Locally Linear Embedding).

2.1 Mathematical Perspective

Given a set of observations $\mathbf{X} = \{x_i\}_{i=1}^n$ in an ambient space of dimensionality D (where $x_i \in \mathbb{R}^D$), the aim of dimensionality reduction is to recover the outputs $\mathbf{Y} = \{y_i\}_{i=1}^n$ in inherent space d ($d \ll D$ and $y_i \in \mathbb{R}^d$) that best represent the subspace or submanifold contained in the ambient space.

2.2 Principal Components Analysis

Principal Components Analysis (PCA) was first discovered by Pearson in 1901 [9] and was further developed by Hotelling in 1933 [10]. It is perhaps the most widely used dimensionality reduction technique and provides the foundation to many other methods. The principal goal of PCA is to maintain maximal variance between the data points in the low dimensional space and as such it finds the subspace \mathcal{S} within the ambient space that has maximum variance. PCA begins by constructing the zero mean covariance matrix, $\mathbf{C} = cov_{\mathbf{X} - \bar{\mathbf{X}}}$, of \mathbf{X} , before finding the solution to the eigenproblem

$$\mathbf{C}\mathbf{W} = \lambda\mathbf{W} \quad (1)$$

The original data, \mathbf{X} , is then projected onto the top n eigenvectors of \mathbf{W} to give the low dimensional representation \mathbf{Y} .

2.3 Locally Linear Embedding

Locally Linear Embedding (LLE) [11] is one of the more popular non-linear dimensionality reduction techniques. LLE, as the name suggests, aims to preserve the local geometry of the manifold by maintaining local neighborhoods in the high and low dimensional spaces. This is achieved by minimizing the embedding cost function

$$\Psi(\mathbf{Y}) = \sum_{i=1}^n |y_i - \sum_{j=1}^n \mathbf{W}_{ij}y_j|^2 \quad (2)$$

The weights contained in the matrix \mathbf{W} will have been previously calculated by minimizing a similar reconstruction error based cost function¹. This can then be minimized by solving an eigenvalue problem whose bottom d eigenvectors

¹ $\epsilon(\mathbf{W}) = \sum_{i=1}^n |x_i - \sum_{j=1}^n \mathbf{W}_{ij}x_j|^2$

provide the set of orthogonal coordinates.

2.4 Intrinsic Dimensionality Estimation

An important, but often under used, tool related to dimensionality reduction is the estimation of the intrinsic dimensionality of the data. The intrinsic dimensionality can be defined as the smallest number of independent parameters that is needed to generate the given data set [12]. When using a classifier it is useful to be able to work in the smallest possible dimensionality as high-dimensional problems lead to more redundant data as well as increased computational complexity. If the intrinsic dimensionality can be correctly estimated then the redundant data can be “stripped-away” and the real (intrinsic) data can speak for itself.

Many techniques exist for estimating intrinsic dimensionality (see [13]) and in this paper we use two methods with widely differing approaches. As dimensionality reduction techniques can be broken up into linear and non-linear, so can intrinsic dimensionality estimators. We have chosen one linear and one non-linear. The first method is closely related to PCA and simply uses the Eigenvalues created from Equation 1 to estimate the dimensionality. By calculating the residuals of the Eigenvalues and finding at which point the biggest “jump” from one value to another occurs the intrinsic dimensionality can be estimated. The second is based on the Geodesic Minimum Spanning Tree of the data [14]. This works by creating a sequence of minimal spanning trees using geodesic distances (obtained by the Isomap [15] algorithm) and uses the overall lengths of the minimum spanning trees to estimate the dimensionality of the manifold.

3 Data

The data comes from features extracted from the whole set of 322 mammograms that form the MIAS database [1, 16]. The data is based on breast tissue density and consists of 322 samples each with 280 features, 10 from morphological characteristics and the remaining 270 from texture information. A fuzzy C-means approach was used to extract two clusters (relating to fatty and dense tissue) from the mammograms. The morphological features were created using relative area of the fatty and dense clusters as well as the first four histogram moments of these clusters. The texture information was derived from co-occurrence matrices [17]. Each of these 322 mammograms have been assigned to a BI-RADS risk assessment class by an expert radiologist [1].

4 Methodology

The first step in this experiment was to obtain classification results using the raw high-dimensional data. A k -fold cross validation technique was employed throughout this experiment. The data was partitioned into two sets: 1 for training the classifier and 1 for testing. The size of each fold was 14 samples. This meant that for each stage of the cross validation experiment 14 of the 322 samples were used for testing the classifier while the remaining 308 were used for training purposes. The average over each of these folds was then used as the high-dimensional result. A simple k -nearest neighbor classifier [18] was used throughout this experiment with the results being averaged over a range of $2 \leq k \leq 30$. The results are averaged so as to try to factor out any effects the classifier parameters might have on the results. We try to solely look at the effects that the dimensionality reduction techniques have. More advanced classifiers could have been used (such as SVM, C4.5 and Bayesian [19]) but the use of these algorithms would have made factoring out parameter effects more difficult.

The outcome of a dimensionality reduction technique is heavily affected by the choice of parameters. So one of the key steps needed when using dimensionality reduction is finding the optimal parameter set. Without this step you run the risk of performing dimensionality reduction at sub-optimal settings, leading to worse classification results. With this in mind a simple parameter search can be used to find the optimal parameters for each technique. For PCA where the only parameter is the target dimensionality the search is straight forward, we simply run the algorithm over a range of dimensions ($1 \leq d \leq 28$). When using LLE the neighborhood size (k) must be specified. So the algorithm was run multiple times over a range of values for k ($2 \leq k \leq 30$) and the optimal value was recorded and used.

Once the optimal parameters have been found the data can then be classified. The results can then be compared against those created in high-dimensional space to see if an improvement occurs. The optimal dimensionality found from the parameter search can also be compared against the estimated intrinsic dimensionality to see if the two do actually coincide.

Ambient Space ($\kappa = 0.50$; $A_c = 56\%$)					PCA + k -NN ($\kappa = 0.57$; $A_c = 63\%$)				
	B-I	B-II	B-III	B-IV		B-I	B-II	B-III	B-IV
B-I	67	30	13	2	B-I	63	20	5	2
B-II	16	62	35	6	B-II	16	67	29	4
B-III	4	11	40	19	B-III	8	16	55	14
B-IV	0	0	7	10	B-IV	0	0	6	17
	62%	62%	60%	46%		72%	65%	58%	46%

LLE + k -NN ($\kappa = 0.53$; $A_c = 59\%$)				
	B-I	B-II	B-III	B-IV
B-I	66	18	2	2
B-II	16	62	36	2
B-III	5	23	50	20
B-IV	0	0	7	13
	76%	60%	53%	35%

Table 1. Confusion Matrices for classification of MIAS database using different dimensionality reduction techniques with optimal parameter sets. The results from classification in high-dimensional ambient space are also shown.

5 Results

The results of the experiments are shown in Table 5 with optimal parameters found to be $PCA(d = 4)$ and $LLE(d = 10, k = 17)$. As well as the confusion matrices the kappa co-efficient and classification accuracy of each experiment is also displayed. The kappa coefficient is a measure of agreement, beyond chance, between the actual results and the predicted results. As can be seen the use of dimensionality reduction improves classification performance over classification in high-dimensional space. PCA gives the best performance with an increase of classification accuracy of 7%. LLE yields an increase of 3%. When examining the kappa co-efficient again PCA yields the biggest improvement signifying that it retains more important aspects of the data between the high and the low-dimensional space. Even though PCA is only a linear technique it still outperforms LLE. One reason for this could be that LLE simply fails to find any meaningful manifold in the high-dimensional space, and so picks up a sub-optimal noisy manifold. Local techniques tend to over fit the manifold and do not necessarily find the global structure of the data.

The graph in Figure 6 shows the classification accuracy of PCA and LLE across a range of dimensions. What is immediately noticeable is the fact that PCA performs best at $d = 4$. After this point there is no noticeable change in the classifier’s accuracy showing that the data can be well expressed using only 4 dimensions. This correlates with the outcome of the intrinsic dimensionality estimators, both of which estimated the intrinsic dimensionality at 4-dimensions. This gives weight to the fact that optimal classification performance occurs when classifying in the intrinsic-dimensional space. LLE’s optimal performance occurs at $d = 10$. The reason for this could be related to the fact that LLE can’t find the manifold on which the data lies. The best estimate of the manifold it can find occurs when reducing to 10-dimensions.

6 Conclusions & Future Work

At the beginning of this paper we posed the questions of how well a simple classification algorithm performs in a reduced dimensional space, in particular the apparent intrinsic dimension of the data. From the results shown in Section 5 we can see that when using PCA on feature data extracted from mammographic images the best classification performance occurs when working in the estimated intrinsic dimension. There is also a noticeable increase in the classification accuracy and kappa co-efficient when using either PCA or LLE. This shows that there are benefits to using a classifier in the estimated intrinsic dimension. We intend to extend this work to show how using more advanced classifiers can yield a greater improvement to the classification accuracy.

A further extension to this work would be to look at how different dimensionality reduction techniques affect the classification performance. In this paper we have only focused on two techniques but other methods may be able to pick up more significant aspects (such as topological features) from the data.

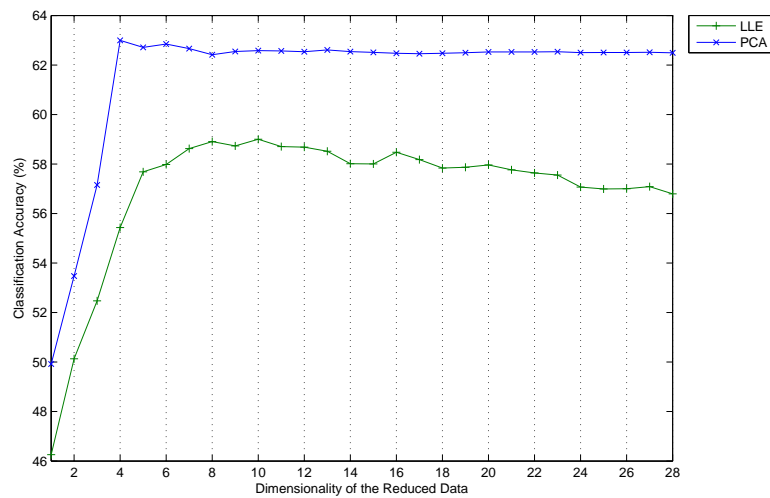


Figure 2. A graph of classification accuracy against dimensionality of reduced data. The optimum of each is PCA($d = 4$) and LLE($d = 10$). The estimated intrinsic dimensionality from both GMST and Eigenvalues was 4. The results from LLE were obtained using $k = 17$.

References

1. A. Oliver, J. Freixenet, R. Marti et al. "A novel breast tissue density classification methodology." *IEEE Transactions on Information Technology in Biomedicine* **12**(1), pp. 55–65, 2008.
2. C. M. Kuzmiak, G. A. Millnamow, B. Qaqish et al. "Comparison of full-field digital mammography to screen-film mammography with respect to diagnostic accuracy of lesion characterization in breast tissue biopsy specimens." *Academic Radiology* **9**, pp. 1378–1382, 2002.
3. T. W. Freer & M. J. Ulissey. "Screening mammography with computer-aided detection: Prospective study of 12860 patients in a community breast center." *Radiology* **220**, pp. 781–786, 2001.
4. Eurostat. "Health statistics atlas on mortality in the european union." *Office for Official Publications of the European Union* 2002.
5. J. N. Wolfe. "Risk for breast cancer development determined by mammographic parenchymal patterns." *Cancer* **37**(5), pp. 2486–2492, 1976.
6. N. Boyd, J. Byng, R. Jong et al. "Quantitative classification of mammographic densities and breast cancer risk: Results from the canadian national breast screening study." *Journal of the National Cancer Institute* (**87**), pp. 670–675, 1995.
7. American College of Radiology. *Illustrated Breast Imaging Reporting and Data System BIRADS*. American College of Radiology, third edition, 1998.
8. L. Tabár, T. Tot & P. B. Dean. *Breast Cancer: The Art And Science of Early Detection with Mammography: Perception, Interpretation, Histopathologic Correlation*. Georg Thieme Verlag, first edition, December 2004.
9. K. Pearson. "On lines and planes of closest fit to systems of points in space." *Philosophical Magazine* **2**, pp. 559–572, 1901.
10. H. Hotelling. "Analysis of a complex of statistical variables into principal components." *Journal of Educational Psychology* **24**, pp. 417–441, 1933.
11. S. T. Roweis & L. K. Saul. "Nonlinear dimensionality reduction by locally linear embedding." *Science* **290**, pp. 2323–2326, 2000.
12. P. J. Verwee & R. P. W. Duin. "An evaluation of intrinsic dimensionality estimators." *IEEE Transactions on Pattern Analysis and Machine Intelligence* **17**(1), pp. 81–86, 1995.
13. J. A. Lee & M. Verleysen. *Nonlinear Dimensionality Reduction*. Springer, 2007.
14. J. A. Costa & A. O. Hero. "Geodesic entropic graphs for dimension and entropy estimation in manifold learning." *IEEE Transactions on Signal Processing* **52**(8), pp. 2210–2221, 2004.
15. J. B. Tenenbaum, V. de Silva & J. C. Langford. "A global geometric framework for nonlinear dimensionality reduction." *Science* **290**, pp. 2319–2322, 2000.
16. J. Suckling, P. J. D. Dance et al. "The mammographic images analysis society digital mammogram database." In *Digital Mammography*, pp. 375–378. 1994.
17. R. M. Haralick, K. S. Shanmugan & I. Dunstein. "Textual features for image classification." *IEEE Transactions on Systems, Man and Cybernetics* **SMC-3**(6), pp. 610–621, 1973.
18. B. V. Dasarthy. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. IEEE Computer Society Press, 1990.
19. S. Theodoridis & K. Koutroumbas. *Pattern Recognition*. Academic Press, third edition, 2006.