Detecting Breast Cancer with Supervised Learning Techniques

Matt Dietrich
50981794
University of British Columbia
dietrich@alumni.ubc.ca

Arun Rajendran
86611860
University of British Columbia
arun95@math.ubc.ca

Abstract

Because many different machine learning approaches can be applied to the field of healthcare, it is important to identify which techniques perform well on a specific task. We target the task of breast tissue tumour classification and compare the performance of several supervised learning classifiers on cancer detection. As a supplementary approach, we also explore unsupervised dimension reduction techniques to visualize the dataset and search for distinguishable clustering. Although each tested classifier performs well, we observe that random forest classifiers yield the lowest test error on average and therefore seem to be best suited to the tumour classification problem.

1 Introduction

Healthcare is a critical pillar of society that continues to drive and leverage advancements in technology. As the availability of data improves, machine learning is increasingly applied to many challenges in the field. The use of machine learning is justified for several reasons, such as the potential for faster medical diagnoses with reduced human error.

Given the plethora of machine learning algorithms available, it is important to understand which techniques are best suited to a specific healthcare related task (such as the diagnosis of cancer). Because early detection provides more time for the delivery of potentially life-saving treatment, it is especially critical to identify and apply high performing classification methods.

In this project, we focus on the classification of tumours as malignant (cancerous) or benign (non-cancerous). We evaluate the performance of several supervised learning techniques in classifying breast tissue tumours in the Wisconsin breast cancer diagnostic dataset retrieved from the UCI Machine Learning repository [2]. The classifiers we consider are random forests, k-nearest neighbours, naive Bayes, support vector machines, and neural networks.

Our primary contribution is the following: we observe that each of the tested classifiers performs fairly well after hyperparameter tuning, with errors below 10%. The random forest classifier yields the lowest test error on the dataset and therefore seems to be the best suited classifier for this task.

As a complement to the classifier comparison, we also explore unsupervised dimensionality reduction similar to Jamieson et al. [3]. We plot the two-dimensional results of principal component analysis, multi-dimensional scaling, ISOMAP, and t-distributed stochastic neighbour embedding. Our visualizations show distinguishable clustering of the data, with several mixed regions. We conclude therefore that unsupervised dimensionality reduction techniques could potentially lend support to a supervised learning classification decision but are unlikely to be sufficient on their own.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

2 Related Work

Polat and Guines [6] conducted breast cancer diagnosis using least square support vector machine (LS-SVM) classifier algorithm and showed that it achieves close to 98% accuracy. Jose et al. [4] evaluated the performance of several statistical and machine learning imputation methods that were used to predict recurrence in patients in an extensive real breast cancer data set. Kourou et al. [5] presented a review of recent machine learning approaches employed in the modeling of cancer progression. Akay [1] proposed breast cancer diagnosis based on a SVM-based method combined with feature selection and showed that it has highest classification accuracy (99.51%) when the model contains five features.

Wolberg et al. [8] develoed an interactive computer system that evaluated and diagnosed breast cancer based on cytologic features derived directly from a digital scan of fine-needle aspirate (FNA) slides. They trained their system on data provided by 569 patients and used samples given by additional 54 consecutive, new patients to test the system. They have shown that the projected prospective accuracy of the system estimated by tenfold cross validation was 97% and the actual accuracy on test dataset consisting of 54 new samples (36 benign, 1 atypia, and 17 malignant) was 100%. They also stated that the digital image analysis coupled with machine learning techniques will improve diagnostic accuracy of breast fine needle aspirates.

These works indicate that high levels of accuracy can be achieved using machine learning algorithms. Most of the works have focused on one particular algorithm and have presented their validation and test results for this algorithm. In our project, we present a comparative study of different supervised machine learning techniques and identify the best algorithm that is suitable for the breast cancer dataset. We believe that this step has to be done before hyperparameter tuning and cross-validation to achieve optimal accuracy gains.

3 Methods

3.1 The Dataset

Our analysis focuses on the Wisconsin breast cancer diagnostic dataset retrieved from the UCI Machine Learning repository [2], originally used in previous work [7]. The features in the dataset are calculated from microscope images of breast tissue masses. The following features are included for each of the 569 examples:

- ID number
- Diagnosis, where M = malignant (cancerous) and B = benign (noncancerous)
- Ten numerical features are computed for each cell nucleus, described in further detail in the original paper [7]:
  - Radius (defined as the mean of distances from center to points on the perimeter)
  - Texture (defined as the standard deviation of gray-scale values)
  - Perimeter (defined as the distance around the shape)
  - Area (defined as the size of the shape interior)
  - Smoothness (defined as the variation in radius lengths)
  - Compactness (defined as the perimeter squared divided by the area)
  - Concavity (defined as the magnitude of concave portions of the contour)
  - Concave points (defined as the number of concave portions of the contour)
  - Symmetry (defined as the length difference between lines perpendicular to the longest chord in the cell)
  - Fractal dimension (defined as the rate of change of measured perimeter as the ruler size increases)

For each of the above ten features, the dataset includes a mean, standard error, and a maximum value for each example. This provides a total of 30 features.

In our experiments, we discard the ID number as an irrelevant feature and use the diagnosis as our class label. We map malignant (cancerous) to 1 and benign (noncancerous) to 0.

Table 1: Hyperparameters tuned for each supervised learning classifier

Classifier	Hyperparameters Tuned
Random forests	number of estimators
k-nearest neighbours	number of neighbours
Gaussian naive Bayes	variance smoothing
Linear support vector machines	C (error term penalty)
Neural networks	hidden layer sizes [1]
	alpha (L2 regularization parameter)

We believe machine learning techniques can help with this dataset because it includes numerical features used by medical practitioners to diagnose tumours and was proven to yield a high performing decision tree classifier in previous work [7].

3.2 Supervised Learning Classifiers

We conduct our main performance evaluation on a variety of common machine learning techniques. Specifically, we apply random forests, k-nearest neighbours, naive Bayes, linear support vector machines, and neural networks to the dataset. We survey a broad selection of supervised learning classifiers because of our goal to discover the best performing techniques for the tissue diagnosis problem.

3.3 Unsupervised Dimension Reduction Techniques

Secondarily, we use unsupervised dimension reduction techniques to provide complementary analysis. We apply principal component analysis, multi-dimensional scaling, ISOMAP, as well as t-distributed stochastic neighbour embedding to the entire dataset. Our intention here is to visualize the dataset in two dimensions to look for any observable clustering. Again, we choose to apply several techniques in a broad survey to search for the most useful visualization for this specific application.

4 Experiments and Evaluation

4.1 Supervised Learning Classifier Comparison

Our primary experiment consists of the performance evaluation of several supervised learning techniques on the breast cancer dataset.

We begin by randomizing the order of the 569 examples and split the dataset so that we have 80% of the examples for training and 20% of the examples set aside for testing. Next, we tune a selection of hyperparameters for each model using grid search with 5-fold cross-validation. Hyperparameters tuned are summarized in Table 1.

We repeat the experiment (including randomizing the dataset order) 10 times and take the mean test error for each classifier to limit irregularities arising from the random test set selection. Our intention is to identify the method or methods that perform the best for this particular application. Average classification error for each classifier is summarized in Figure 1.

4.2 Unsupervised Dimension Reduction Visualization

Our secondary analsysis consists of unsupervised dimesion reduction of the dataset. We proceed to directly transform the full 569 examples in dataset into two dimensions. Our intention is to look for any observable clustering in the dataset that could support classification. Results of these two-dimensional visualizations are shown in Figure 2.

[1] We consider only one hidden layer in our neural network analysis.

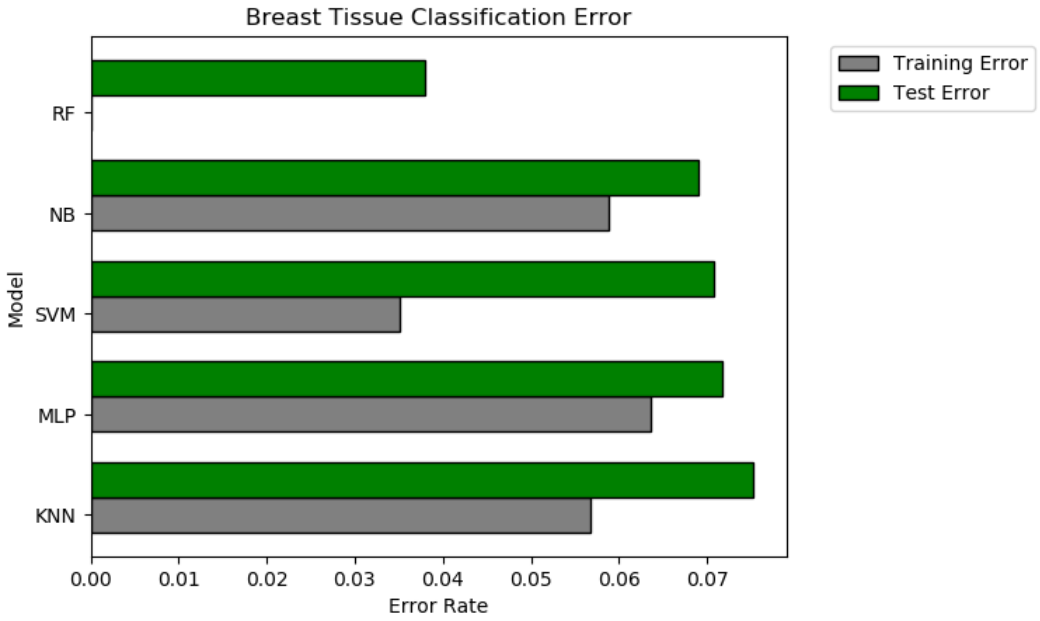
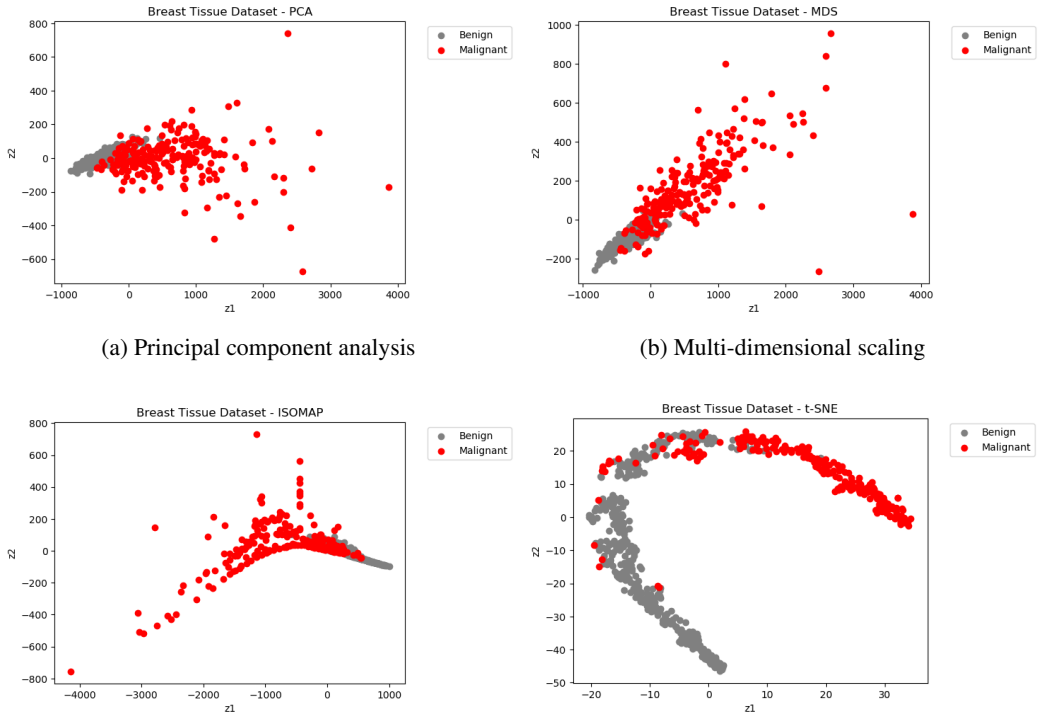


Figure 1: Average classification error after hyperparameter tuning. Model name abbreviations are the following: RF = random forest; NB = Gaussian naive Bayes; SVM = linear support vector machine; MLP = multilayer perceptron (neural network); KNN = k-nearest neighbours.



(a) Principal component analysis

(b) Multi-dimensional scaling

(c) ISOMAP, n_neighbors = 5

(d) T-distributed stochastic neighbour embedding

Figure 2: Visualizations obtained from unsupervised dimension reduction techniques

5 Discussion and Future Work

Our supervised learning results show that each of the surveyed techniques perform well on the breast cancer dataset. We observe that each model achieves an average test error of less than 10%. We also conclude that random forest classifiers perform the best on average for this dataset, with an average test error of approximately 4%.

Among our dimension reduction visualizations, t-distributed stochastic neighbour embedding arguably provides the most uniform clustering of the dataset. However, we still see significant regions of the visualization with a mix of malignant (cancerous) and benign (noncancerous) class labels. We therefore conclude that unsupervised dimension reduction is insufficient on its own to accurately classify breast tissue tumours, although it could lend support to a supervised classifier.

An important limitation of this project is the fact that only one dataset was examined and that it is limited to 569 examples. Future efforts could evaluate more datasets and perform more extensive hyperparameter tuning. While this project included a broad survey of machine learning techniques, it would be useful to deepen the investigation to include additional variations, such as kernel support vector machines. This would yield more definitive conclusions given the more extensive nature of the analysis.

Furthermore, increasingly complex tasks could be explored. It would be interesting to target datasets of tumour images directly with deep learning techniques such as convolutional neural networks. This would further streamline the diagnosis procedure and eliminate the need for a healthcare practitioner to manually compute the features from each image, as is the case with the dataset we used.

Despite these challenges, our work here affirms the usefulness of random forest classifiers for breast tissue classification amidst a wide array of machine learning algorithms. Medical software application developers are encouraged to consider random forests when applying machine learning in their products.

References

[1] Mehmet Fatih Akay. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications*, 36(2, Part 2):3240 – 3247, 2009. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2008.01.009. URL http://www.sciencedirect.com/science/article/pii/S0957417408000912.

[2] Dua Dheeru and Efi Karra Taniskidou. UCI Machine Learning Repository, 2017. URL http://archive.ics.uci.edu/ml.

[3] Andrew R Jamieson, Maryellen L Giger, Karen Drukker, Hui Li, Yading Yuan, and Neha Bhooshan. Exploring nonlinear feature space dimension reduction and data representation in breast cadx with laplacian eigenmaps and t-sne. *Medical physics*, 37(1):339–351, 01 2010. doi: 10.1118/1.3267037. URL https://www.ncbi.nlm.nih.gov/pubmed/20175497.

[4] José M. Jerez, Ignacio Molina, Pedro J. García-Laencina, Emilio Alba, Nuria Ribelles, Miguel Martín, and Leonardo Franco. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, 50(2):105 – 115, 2010. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2010.05.002. URL http://www.sciencedirect.com/science/article/pii/S0933365710000679.

[5] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8 – 17, 2015. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2014.11.005. URL http://www.sciencedirect.com/science/article/pii/S2001037014000464.

[6] Kemal Polat and Salih Güneş. Breast cancer diagnosis using least square support vector machine. *Digital Signal Processing*, 17(4):694 – 701, 2007. ISSN 1051-2004. doi: https://doi.org/10.1016/j.dsp.2006.10.008. URL http://www.sciencedirect.com/science/article/pii/S1051200406001461.

[7] O. L. Mangasarian W. Nick Street, W. H. Wolberg. Nuclear feature extraction for breast tumor diagnosis, 1993. URL https://doi.org/10.1117/12.148698.

[8] William H. Wolberg, W.Nick Street, and O.L. Mangasarian. Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. *Cancer Letters*, 77(2):163 – 171, 1994. ISSN 0304-3835. doi: https://doi.org/10.1016/0304-3835(94)90099-X. URL http://www.sciencedirect.com/science/article/pii/030438359490099X. Computer applications for early detection and staging of cancer.