Effects of Large Constituent Size in Variable Neural Ensemble Classifier for Breast Mass Classification

Peter Mc Leod and Brijesh Verma

Central Queensland University
Bruce Highway, North Rockhampton QLD 4702
mcleod.ptr@gmail.com, b.verma@cqu.edu.au

Abstract. This paper proposes a novel ensemble technique for mass classification in digital mammograms by varying the number of hidden units to create diverse candidates. The effects of adding more networks to the ensemble are evaluated on a mammographic database and the results are presented. A classification accuracy of ninety nine percent is achieved.

Keywords: Ensemble classifiers, neural networks, digital mammography.

1 Introduction

Breast cancer has increased in prevalence. The aetiology is unknown and a cure does not seem likely [1]. Research has progressed in relation to treatment but this relies on an accurate diagnosis however 11-25% of cancers are missed [2]. Reasons include distortion of the breast, occlusions with surrounding tissue, low mammogram contrast and even talc on the breast. The rate of breast cancer is low with three to four malignancies in a thousand [3]. A high volume of mammograms means that skill levels, complacency and fatigue can impact on radiologists. An estimated 35% of biopsies are not required [4] resulting in stress to patients and increased load on the health system. Despite this digital mammography is the diagnostic tool of choice due to wide availability, low cost and its non-invasive nature. Mechanisms such as a second radiologist to rescreen mammograms have been shown to improve the classification rate and reduce misdiagnosis. The cost and volume of mammograms makes this ineffective. Mechanisms including Computer Assisted Diagnostic (CAD) systems to act as an adjunct to radiologists have been suggested however variable classification accuracy has been a problem. This has been researched for around 40 years and arguably neural networks have demonstrated good capabilities. Techniques used to improve this situation include the use of many classifiers in a voting arrangement (ensemble). This research aims to create an accurate ensemble classifier.

This paper is broken into several sections with section 2 covering the research background. Section 3 details the proposed methodology while section 4 details the results. Discussions and analysis are in section 5 while section 6 details our conclusions and future research.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 525–532, 2013.

2 Background

Costa, Campos and Barros [5] used efficient coding based on Independent Component Analysis (ICA) achieving an accuracy of 90.07% on 5090 anomalies from the Digital Database of Screening Mammography (DDSM). They developed a compact code based on a statistics pattern ensemble to reduce redundancy with minimal loss of information. The data is transformed by linear functions generating an estimate of independent components. They used 41 components performing better than Principal Component Analysis (87.28% with 39 principal components) and Gabor Filter (85.28%). Luo and Cheng [6] used a bagged Decision Tree (DT) to gain an accuracy of 83.4% on mass anomalies. They utilized a DT and Support Vector Machine (SVM) Sequential Minimal Optimization. Mass anomalies from the University of California at Irvine (UCI) were classified using feature selection techniques to reduce the BI-RADS® input features from five to four. Mass margin was the most important feature. Their ensemble was more effective than using a single classifier. Yoon [7] achieved an area under the ROC curve of 0.94315 Az on a DDSM mass dataset with a boosted SVM ensemble together with fivefold cross validation to select the most appropriate features. Verma et al. used a partitioning mechanism for training of a classifier with direct output weight calculation by least squares (modified gram-schmidt) resulted in the creation of a Soft Clustered Neural Network (SCNN) [8] with 94% classification accuracy on mass anomalies from the DDSM. This technique removed those clusters that did not contribute to a class assignment in order to create a better decision boundary. The least squares technique does not suffer from local minima. Techniques of identifying sub-populations (soft-clusters) for the benign and malignant patterns to reduce class variability and increase classification accuracy on a neural network have also been used. This approach was known as Soft Clustered Based Direct Learning (SCBDL) [9] and achieved a classification accuracy of 97.5% on a dataset from the DDSM. Another approach used a SVM classifier with a genetic algorithm to select the classifier features [3]. This research attempted to test a new feature selection technique on a DDSM dataset with an accuracy of 89% being achieved. Other researchers examined mechanisms to create ensemble classifier; determining that 3-5 different classifiers were optimal taking into account diversity and variability [10].

3 Proposed Methodology

Neural networks are interconnected processing systems where each connection responds to input and the resultant outputs from the interconnected units (neurons) are aggregated to form a decision. Neural networks are capable of reaching a decision by the weights that interconnect the layers of neurons in the network. Through training knowledge of how to reach a decision is built into the weights.

Researchers examined the issue of obtaining the best possible configuration for neural networks with the selection of the best number neurons of being an area that was not investigated fully as the performance improvement was low. Investigations

utilized only a small number of neurons in the hidden layer [11]. Others noted that too high a number was associated with overtraining [12, 13]. Diversity (or disagreement) is a key concept for the creation of ensembles. Diversity is the concept that a classifier is right more often than not; however when compared to another classifier its decision boundary is sufficiently different that it does not misclassify the same patterns. Combining the results of diverse classifiers should yield a result better than any single classifier. The proposed technique creates diverse classifiers to build an ensemble, as depicted in Figure 1. A detailed discussion of the system follows.

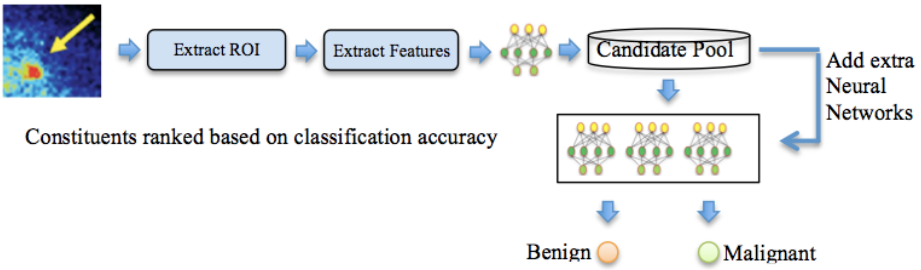

Fig. 1. Proposed variable neuron based ensemble technique

3.1 Mammograms

The mammographic images for this research (100 malignant and 100 benign mass anomalies) are from the DDSM. This is one of the largest publicly available benchmark databases with 2600+ images. The anomalies are fully annotated with case information, cancer has been proven with biopsy and patients have been followed for a number of years to ensure that benign cases are indeed benign. Images are stored using a lossless compression algorithm, ensuring a high quality dataset.

3.2 Region of Interest

Mammographic images are large images to process and a diagnostic process is only concerned with making a diagnosis about a small area (anomaly). To conserve computational resources (memory and cpu capacity) only the Region Of Interest (ROI) (anomaly) is examined by extracting a boundary around the anomaly. The DDSM has a chain code for this process. Extracting the ROI does not attempt to classify an anomaly.

3.3 Feature Extraction

Once an anomaly has been extracted it is necessary to obtain the features that are used to form a decision as to whether it is malignant or benign. Breast masses are not easy to classify and no one feature can be used so multiple features are used. The features utilized in this research are based on the Breast Imaging Reporting and Data System

(BI-RADS®) as well as patient age and a subtlety value. BI-RADS® features have a positive predictive capacity for predicting mass malignancy [1, 14, 15]. The shape, density and mass margins are morphological features, which are utilized by radiologists. In some cases the pathology cannot be confirmed until histological samples are obtained and examined through biopsy. The difficult nature of performing a classification without a biopsy has been shown with the benign rate of biopsies being 65-90% [14]. Utilizing a feature set rather than a single feature increases classification accuracy however too many can reduce accuracy [16]. The features used in this research are patient age [17] (more aggressive tumors in younger patients), anomaly density (if the same density as surrounding tissue then hard to detect), shape (spiculated margins infer invasive tumors), margin (indistinct margins indicate harder to find and potentially more aggressive), subtlety (how hard is it to find) and assessment rank (a ranking of likely seriousness).

3.4 Network Training

A large number of neural networks are created by varying the number of neurons in a single hidden layer (from 2 to 1001) creating a large number of candidates for the ensemble. Changing the network parameters results in different weights between the layers, creating diverse classifiers. The candidates are created with the following parameters. Ten-fold cross validation is incorporated during training and testing. A Root Mean Square (RMS) error of 0.001 or (a maximum of 3000 iterations) is used for the stopping criteria. A learning rate of 0.05, momentum of 0.7 with six input neurons and two output neurons is used. Hyperbolic tangent sigmoid (tansig) is the transfer function between the layers with the system implemented in MATLAB™.

3.5 Ensemble Creation

The ensemble is created from the candidate pool with candidates ranked according to classification accuracy, which is the only inclusion mechanism. The first ensemble created is comprised of three neural networks. It is trained, tested and then another neural network is added with the process repeating to create a new ensemble of four neural networks (this is represented by the arrow in Figure 1.) This continues until an ensemble composed of 202 candidates is created. An upper bound of 202 is chosen to determine the effect of a large number of constituents (200 ensembles in total).

3.6 Classification and Fusion

Individual classifier results in the ensemble are fused together to form a classification using the majority vote algorithm, as it is one of the simplest but effective fusion mechanisms. In the event of a tie the smallest output value is chosen representing a malignant pattern. A false diagnosis for a malignant condition would be more severe than a false classification for a benign condition.

4 Experiments and Results

Experiments are conducted to create a candidate pool (one thousand) of back propagation neural network classifiers that had a different number of hidden units in the single hidden layer. It was hypothesized that this would be diverse enough to create an ensemble classifier with good accuracy.

Table 1. Performance of neural network on breast mass dataset (candidate classifiers)

Hidden Units	True Positive	False Negative	Accuracy
823	87	88	87.5
242	86	87	86.5
400	87	86	86.5
592	79	83	81.0
1000	82	78	80.0
78	69	81	75.0

Table 2. Performance of ensemble network on breast mass dataset

Constituents	Configuration	Accuracy (%)
3	823,242,400	95.0
4	823,242,400,24	92.5
10	823,242,400,24,262,302,404,657,5,15	97.5
100	823,242,400,24,262,302,404,657,5,15,32,268,281,292, 309,494,550,31,43,50,75,158,165,183,209,224,349,355 ,356,398,416,426,436,443,466,473,622,639,659,661,67 8,749,903,904,925,38,59,68,79,116,146,168,175,204,2 18,223,232,233,235,243,246,254,277,297,304,305,325, 350,352,366,388,395,417,427,444,459,471,493,500,53 7,546,556,583,612,664,682,739,753,842,866,870,887,9 30,957,999,14,30,37,95,103	98.5
127	823,242,400,24,262,302,404,657,5,15,32,268,281,292, 309,494,550,31,43,50,75,158,165,183,209,224,349,355 ,356,398,416,426,436,443,466,473,622,639,659,661,67 8,749,903,904,925,38,59,68,79,116,146,168,175,204,2 18,223,232,233,235,243,246,254,277,297,304,305,325, 350,352,366,388,395,417,427,444,459,471,493,500,53 7,546,556,583,612,664,682,739,753,842,866,870,887,9 30,957,999,14,30,37,95,103,104,138,140,166,171,174, 187,188,202,212,221,252,259.282,288,312,340,367,36 8,384,391,421,438,470,510,551,573	99.0

Our literature review indicates that limited research into the creation of diverse networks by varying the number of hidden units in the hidden layer has been undertaken. The accuracy of the candidate networks ranged from 75% to 87.5%.

The candidate networks are ranked in descending order based on performance. The highest performers are selected for inclusion in the ensemble. Table 1 provides a summary of the classification accuracy achieved. Table 2 shows a subset of the accuracy achieved from the ensemble networks. Combining the best performing candidate networks created the ensemble.

5 Discussion

The results demonstrate that only a few candidates are needed to improve classification accuracy although this is variable in the early stages. To substantiate that an improvement in classification accuracy is achieved over the neural network an ANOVA analysis of variance is performed to see if the improvement is statistically significant (Table 3) using a 5% confidence level.

Table 3. ANOVA analysis summary

Groups	Count	Sum	Average	Variance
MLP	100	8485	84.85	0.335859
Ensemble	100	9815	98.15	0.063131

Table 4. ANOVA analysis details

	SS	*df*	*MS*	*F*	*P-value*	*F crit*
Between	8844.5	1	8844.5	44334.46	8.0331E-235	3.888853

In Table 4, the p-value is significantly below the confidence level confirming the improvement is statistically significant. The variance indicates the ensemble is more stable than the MLP network. Graphing accuracy of the ensemble against the number of classifiers shows a trend of higher accuracy as more classifiers are added. This levels off after around twenty classifiers (Figure 2). The highest classification accuracy of 99% is reached with 76 and 127 constituents. Stratification of the results is performed to determine the population variance as more classifiers are added.

Table 5. Ensemble variance, median and mode for ensemble groupings

No. Of Constituents	Variance	Median	Mode
3-12	3.10000	96.25	97.00
13-22	0.46944	97.25	97.50
53-62	0.19167	98.00	98.00
63-72	0.05556	98.00	97.50
163-172	0.10000	98.00	98.00
173-182	0.02500	98.00	98.00
183-192	0.04444	98.00	98.00
193-202	0.06944	97.75	98.00

A grouping of ten ensembles is chosen for each population in order to examine the changes of adding more classifiers. A subset of results is shown in Table 5. Variance tapers off as more classifiers are added (63-72 classifiers) then increases and tapers off again. In order to evaluate the performance of the proposed system it is necessary to compare its performance against that achieved by other researchers (Table 6).

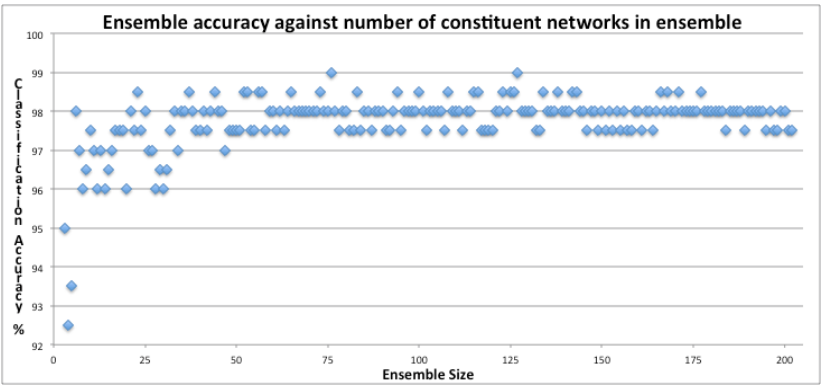

Fig. 2. Ensemble accuracy versus number of constituent classifiers

Table 6. Accuracy obtained by current research in comparison to proposed approach

Luo and Cheng [6]	Elfarra et al. [3]	Costa et al. [5]	Verma et al. [9]	Proposed
83.40%	89.00%	90.07%	97.5%	99.00%

6 Conclusions and Future Research

The variable neuronal ensemble has resulted in a high classification rate (99%) on the test dataset. This is high in comparison to other techniques. A disadvantage is that a high number of candidate networks are required to achieve high classification accuracy. It is noted that after a point adding more classifiers does not increase accuracy. This research uses a simplistic inclusion mechanism of accuracy. Our future research will use a multi-objective genetic algorithm with both diversity and accuracy.

References

1. Orel, S., Kay, N., Reynolds, C., Sullivan, D.: BI-RADS categorization as a predictor of malignancy. Radiology 211, 845–850 (1999)
2. Goergen, S., Evans, J., Cohen, G., Macmillan, J.: Characteristics of breast carcinomas missed by screening radiologists. Radiology 204, 131–135 (1997)
3. Elfarra, B.K., Abuhaiba, I.S.I.: New feature extraction method for mammogram computer aided diagnosis. International Journal of Signal Processing, Image Processing and Pattern Recognition 6, 13–36 (2013)

4. Isaac, L., Richard, L., Shalom, B., Yossi, S., Philippe, B., Fanny, S.: Computerized classification can reduce unnecessary biopsies in bi-rads category 4a lesions. In: Astley, S.M., Brady, M., Rose, C., Zwiggelaar, R. (eds.) IWDM 2006. LNCS, vol. 4046, pp. 76–83. Springer, Heidelberg (2006)
5. Costa, D., Campos, L., Allan, B.: Classification of breast tissue in mammograms using efficient coding. Biomedical Engineering OnLine 10 (2011)
6. Luo, S., Cheng, B.: Diagnosing breast masses in digital mammography using feature selection and ensemble methods. Journal of Medical Systems 36, 569–577 (2010)
7. Yoon, S., Kim, S.: AdaBoost-based multiple SVM-RFE for classification of mammograms in DDSM. In: IEEE International Conference on BioInformatics and Biomedicine (BIBMW 2008), Philadelphia, PA (2008)
8. Verma, B., McLeod, P., Klevansky, A.: A novel soft cluster neural network for the classification of suspicious areas in digital mammograms. Pattern Recognition 42, 1845–1852 (2009)
9. Verma, B., McLeod, P., Klevansky, A.: Classification of benign and malignant patterns in digital mammograms for the diagnosis of breast cancer. Expert Systems with Applications 37, 3344–3351 (2010)
10. West, D., Mangiameli, P., Rampal, R., West, V.: Ensemble strategies for a medical diagnostic decision support system: a breast cancer diagnosis application. European Journal of Operational Research 162, 532–551 (2005)
11. Partridge, D., Yates, W.: Engineering multiversion neural-net systems. Neural Computing 8, 869–893 (1996)
12. Hunter, D., Yu, H., Pukish, M.S.I., Kolbusz, J., Wiliamowski, B.M.: Selection of proper neural network sizes and architectures - A comparative study. IEEE Transaction on Industrial Informatics 8, 228–240 (2012)
13. Lawrence, S., Giles, C.: Overfitting and neural networks: conjugate gradient and backpropagation. In: International Joint Conference on Neural Networks, Como, Italy, pp. 114–119 (2000)
14. Vadivel, A., Surendiran, B.: A fuzzy rule-based approach for characterization of mammogram masses into BI-RADS shape categories. Computers in Biology and Medicine 43, 259–267 (2013)
15. Mu, T., Nandi, A., Ranayyan, R.: Classification of breast masses using selected shape, edge-sharpness and texture features with linear and kernel-based classifiers. Journal of Digital Imaging 21, 153–169 (2008)
16. Kim, S., Yoon, S.: Mass lesions classification in digital mammography using optimal subset of BI-RADS and gray level features. In: 6th International Special Topic Conference on Information Technology Applications in Biomedicine, pp. 99–102 (2007)
17. Tabar, L., Fagerberg, G., Chen, H.-H., Duffy, S., Smart, C., Gad, A., et al.: Efficacy of breast cancer screening by age. New results swedish two-county trial. Cancer 75, 2507–2517 (1995)