Copyright © 2010 Institute of Electrical and electronics Engineers, Inc.

All Rights reserved.

Personal use of this material, including one hard copy reproduction, is permitted.

Permission to reprint, republish and/or distribute this material in whole or in part for any other purposes must be obtained from the IEEE.

For information on obtaining permission, send an e-mail message to stds-igr@ieee.org.

By choosing to view this document, you agree to all provisions of the copyright laws protecting it.

Individual documents posted on this site may carry slightly different copyright restrictions.

For specific document information, check the copyright notice at the beginning of each document.

Soft Clustering and Support Vector Machine based Technique for the Classification of Abnormalities in Digital Mammograms

Peter Mc Leod¹, Brijesh Verma², Minyeop Park

[#]*School of Computing Sciences, CQUniversity
Rockhampton QLD 4702, Australia*

¹pmcleod@civica.com.au ²b.verma@cqu.edu.au

Abstract—This paper presents a novel technique which is the amalgamation of a clustering mechanism and a Support Vector Machine classifier. The technique is called Soft Clustering based Support Vector Machine and is designed to provide a fast converging network with good generalization ability leading to an appropriate classification as a benign or malignant class for the classification of suspicious areas in digital mammograms. The proposed technique has been evaluated on a benchmark database. The experimental results and analysis of results are included in this paper.

The rest of this paper is organized as follows. Section two details existing methods and accuracies for the use of intelligent systems for breast cancer diagnosis. The Soft Clustering and Support Vector Classifier methodology is described in section three with experimental results being presented in section four. These results are discussed with a comparative analysis performed in section five and conclusions together with future research directions being addressed in section six.

I. INTRODUCTION

Medical diagnosis based on digital images is one of the greatest challenges for artificially intelligent systems due to a lack of correlation between the diagnostic features and the resultant disease [1]. In many instances even medical experts cannot provide a sufficiently detailed discourse which will elicit a set of rules that can facilitate classification for all but the simplest of conditions. This dilemma is known as the “knowledge acquisition bottleneck”. Techniques such as artificial neural networks and support vector machines have demonstrated their efficacy in this domain as their ability to classify diseases is not dependent on an embodied set of rules and they have an ability to generalize to an unseen population [2-4]. This means that artificially intelligent systems are slowly being incorporated into the clinical medical diagnostic paradigm [1, 5]. Breast Cancer is a leading cause of mortality worldwide, with more than 1.2 million cases diagnosed each year [6] and has been an area which has leant itself to the application of Computer Aided Diagnostic (CAD) systems. The main problem of CAD systems has been the variable classification accuracy, especially the number of false positive classifications [7, 8]. In this research we propose a novel approach in order to classify breast masses as either benign or malignant by incorporating clustering [9] together with a Support Vector Machine (SVM) classifier. This research has been performed on images from the Digital Database of Screening Mammography (DDSM) [10] and has been subsequently analysed and contrasted with existing research in order to evaluate its efficacy.

II. BACKGROUND

Several techniques have been employed to solve the problem of variable classification accuracy for CAD systems in relation to breast cancer diagnosis. Manrique et al. [11] implemented a Radial Basis Function (RBF) network which removed neurons that did not contribute to achieving a good means square error during the training of a neural network. This technique known as Genetic Algorithm Radial Basis Function (GARBF) neural network achieved an 83% classification accuracy (with 83% specificity and 81% sensitivity) from a dataset from the Madrid hospital (containing both mass and micro-calcification anomalies). Halkiotis et al. [12] utilized a Multi-Layer Perceptron (MLP) type neural network to obtain a detection and classification rate of 94.7% with 0.27 false positives per image for micro calcifications based on the MIAS database. Their technique utilized a mathematical morphology mechanism based on the properties of the mammogram. A potential disadvantage of their technique is that it may need to be calibrated based on the mammograms utilized as the optical and spatial density are taken into account for each dataset, thus requiring recalibration.

Meanwhile Georgiou et al. [13] utilized shape features together with the C-Support Vector Machine (C-SVM) to obtain 91.54% classification accuracy on masses. A custom dataset was utilized for their research where they were determining the shape features that resulted in the highest classification accuracies. Varela et al. [14] utilized a strategy for extracting gray level and morphological features from a custom mass dataset for the training of a neural network

which lead to an Az of 0.90. Archarya et al. [15] obtained a sensitivity of 91.67% using an Artificial Neural Network (ANN) and 95% using a Gaussian Mixture Model (GMM) with 93.33% and 96.67% specificity respectively on micro calcifications from the DDSM. Archarya et al. [15] were investigating a comparative approach for the classification of three kinds of mammograms (normal, benign and cancerous). Verma [16] however added an additional neuron for benign and malignant classes to the hidden layer of an MLP style neural network to achieve 100% training accuracy and 94% classification accuracy on masses from the DDSM. Verma also replaced the traditional gradient descent mechanism of the network with the incorporation of least squares to avoid network paralysis.

Table 1 presents an overview of current accuracies for cancer diagnosis utilizing various soft computing techniques.

TABLE I
CLASSIFICATION ACCURACY FOR CAD TECHNIQUES ON DIGITAL MAMMOGRAMS

Author	Dataset	Classifier	Accuracy %
Manrique [11]	Madrid – Micro/Mass	GARBF	83.00
Halkiotis [12]	Mias – Micro	MLP	94.70 (0.27fp/image)
Georgiou [13]	Custom – Mass	C-SVM	91.54
Verma [16]	DDSM - Mass	MLP	94.00

As can be seen neural networks have demonstrated some good accuracies however Support Vector Machines (SVMs) have made in-roads in this field in recent times. The advantages that SVM offer over Multi-Layer Perceptron style Neural Networks are as follows.

- Maximization of the Networks generalization ability especially where training data is sparse and the data is linearly separable.
- It is not prone to local minima as it is a quadratic equation with a solution.
- SVMs have been considered easier to use/train than Neural Networks [17].

However SVMs have also been shown to have some problems in comparison with MLP Style networks such as:

- Not easily applied to multi class (non binary) problems.
- Can exhibit long training times due to the number of variables affecting the quadratic solution thus potentially being memory, CPU and time restrictive.
- Susceptible to the selection of numerous parameters which can impact on the classifiers operation (Kernel, Margin parameter etc). This is also the same type of problem that MLPs suffer from.

This research attempts to overcome some of the problems faced by existing classifiers. In our research the number of input features has been limited to six features which have been shown to provide good classification accuracy [18-20]

while allowing the quadratic solution to be calculated relatively quickly.

III. CLUSTERING BASED SUPPORT VECTOR MACHINE CLASSIFIER

A high level overview of the research methodology employed is detailed in Figure 1 below together with a synopsis of each step involved in the process.

Fig. 1. Overview of the Research Methodology

This research concentrates on the classification of breast mass abnormalities into benign or malignant classes as this has proven to be a major problem in early detection of breast cancer due to the lack of correlation between the input feature and the subsequent classification as a benign or malignant class. The incorporation of the cluster targets generated from clustering into the Support Vector Machine Classifier is an induction step where we are transforming the raw input features of our data into a target representation that is more suited to the classification capabilities of the Support Vector Machine.

The proposed technique involves the determination of target classes separately for each class by clustering the input features into a number of clusters. This is in order to recognize that some classes may fall into more than a binary classification and that sub-populations may exist.

A. Mammograms - DDSM

The mammographic images utilized for this research are mass type anomalies comprised of malignant and benign cases sourced from the Digital Database of Screening Mammography [10]. The database is available from <http://marathon.csee.usf.edu/Mammography/Database.html>.

This database was chosen as it is publically available, has a large collection (2600+) of high quality images and is the premier choice for images utilized in digital mammography research. The dataset is comprised of 100 malignant and 100 benign samples.

B. Area Extraction

The isolation of an anomalous tissue region reduces the computational expense and time associated with image processing and is a process that is known as area extraction or image segmentation. Images from the DDSM [10] have an “overlay” file which provides a chain code that can be utilized to extract the anomaly together with a surrounding boundary area.

C. Feature Extraction

In mammographic images, textural and gray level spatial features have played a key role in the correct classification of suspicious areas [18-20] and the features utilized for this research have been based on these findings. The six features that have been chosen are four descriptor features from the BI-RADS lexicon together with patient age and a subtlety value [10]. All six features are:

- Density,
- Mass Shape
- Mass Margin
- Abnormality Assessment Rank
- Patient Age
- Subtlety Value

D. Clustering

The process of clustering searches for an optimal partition within a fixed number of clusters by grouping data into d-dimensional feature space with a view to maximizing the similarity within a cluster group as compared to the dissimilar data within other clusters. In this research the k-means clustering algorithm has been utilized due to its simplicity and wide usage within research circles.

The problem with artificially imposing a single hard boundary for the cluster is that it forces the input features into an unrepresentative relationship between the input features and the subsequent malignant or benign class. In order to alleviate this dilemma we can create multiple clusters to take into account the natural variation within each class and provide output targets which assist the SVM classifier by being more linearly separable. This process can be visualised in Figure two where we can see that our Benign Class has been broken down into several benign sub classes that make up the benign class.

This should make each target group more representative as variability due to sub-groups will be reduced and thus with the

appropriate number of clusters should lead to improved accuracy, as the effect of outliers should be minimised. The nature of a benign or malignant class to fall within different clusters when clustered separately substantiates that sub-groupings of these classes exist. These sub classes represent the target output for the SVM classifier as they will produce a more closely grouped classification target and hence a more optimal separating hyperplane describing the sub classes compared to the original class.

Fig. 2. Overview of the Benign Sub Class creation for Benign Masses

E. Training of Support Vector Machine Classifier

The inclusion of clustering has been used by researchers in order to avoid local minima in neural networks and to provide an optimized initialization of weights for a neural network before the training has been initiated. However in this instance instead of using weights from clustering we are using the cluster grouping that has been generated as the target for the SVM classifier.

As such the target class maximizes the similarity within each clustered group and the class to which each clustered group belongs. By extending the usage of incorporating cluster targets to a support vector machine classifier we are providing a more linearly separable input feature set which assists the classifier to appropriately classify the class as either a malignant or benign sub group.

The use of clustering to obtain a better target class should mean that our points in feature space should be more closely grouped and correlated with an output classification which should theoretically lead to a smaller value of the slack variable being required.

Thus an ideal feature for a linear classifier such as SVM would be linearly correlated with the output class. This mechanism of target class selection has been employed for several reasons. Firstly it is known that feature selection can have a profound effect on a classifier's classification accuracy. Conversely selecting the target class so that it is correlated to the input feature should improve classification accuracy. Thus such treatment of the target class can potentially have a profound effect on classification accuracy without attempting to optimize a particular kernel which could prove computationally expensive and may not be able to be generalized to other datasets / problem domains. Thus the incorporation of clustering targets is a computationally efficient and potentially transferrable mechanism for classifier optimization.

The training and testing procedures for the proposed technique based on support vector machine classifier are described below.

Training procedure

- The mammograms from training set are processed and features from suspicious areas are extracted.
- All features from benign class are clustered using k-means into clusters (e.g. 3 clusters/sub classes).
- All features from malignant class are clustered using k-means into clusters (e.g. 3 clusters/sub classes).
- The new training data with input features and output classes (e.g. 3 for benign and 3 for malignant) is created.
- The SVM classifier with new training data is trained.

Test procedure

- The mammograms from test set are processed and features from suspicious areas are extracted.
- The input feature vector is passed to SVM and output is checked.
- If the input feature is from benign class and SVM's output (highest from all SVM outputs) found in previous step is from one of the benign sub-classes (e.g. first 3 benign, last 3 malignant), then it is counted as a correct classification otherwise it is counted as an incorrect classification.
- The above two steps are repeated for each test input feature vector.
- The final accuracy on test data is calculated.

IV. RESULTS

Several rounds of experiments were performed to evaluate the proposed technique. All experiments were performed on the same dataset in order to remove any data centric variance. Unless otherwise stated we used the Weka v3.7 SMO linear classifier [21] with the build logistics model turned on for all experiments performed. The results of the experiments are shown in Table 2.

TABLE II
CLASSIFICATION ACCURACY FOR PROPOSED TECHNIQUE

Kernel	Clusters	Folds	Filter	Accuracy (%)
Normalized PolyKernel	1	2	Standardize	64.00
			Normalize	73.00
		4	Standardize	68.5
			None	71.00
		10	Normalize	72.00
	2	2	None	75.5
			Normalize	81.00
			Standardize	65.5
		4	Standardize	68.00
			None	74.00
		10	Normalize	77.50
	3	2	Standardize	64.50
		4	Standardize	65.00
		6	Standardize	68.50
		10	None	75.00
	5	4	Normalize	76.5
			None	74.50
		10	Normalize	78.00
			None	73.50
PolyKernel	1	2	Standardize	85.00
			Normalize	86.50
		4	Standardize	85.00
			None	85.50
		10	Standardize	86.50
	2	2	None	83.50
			Normalize	85.50
			Standardize	84.50
		4	Standardize	86.00
			None	84.00
	3	2	Standardize	82.00
			Standardize	89.00
			Standardize	91.00
		4	Standardize	89.50
			None	88.50
RBFKernel	1	2	Standardize	86.50
			Normalize	83.50
		4	Standardize	87.00
			None	83.50
		10	Standardize	85.00
	2	2	None	83.50
			Standardize	85.00
			Normalize	85.00
		4	Standardize	84.00
			None	83.50
	3	2	Standardize	84.00
			Standardize	87.50
			None	86.00
		4	Standardize	85.00
			None	85.50
	5	4	Standardize	85.00
			None	85.50
		10	Standardize	85.00

It should be noted that Clusters in Table 2 refers to the number of clusters for each input class (malignant or benign

classes). Therefore the number of classes that the classifier will be classifying is twice this number. The one cluster result represents a standard binary classification dilemma (i.e. one cluster for each class).

V. DISCUSSION

In order to analyse the results that have been obtained we need to compare them with the current research. However at the same time variances between datasets does introduce some variability. As such the analysis of the results obtained is performed both on our prior research [5] utilizing the same dataset and also with current research within the field in order to provide a comprehensive analysis. The highest accuracy achieved was 91% for three clusters using the PolyKernel kernel (and four folds) which is an increase of 4.5% compared to the standard binary classification rate of 86.5% (10 fold) which demonstrates that the proposed clustering based support vector machine classifier can improve the classification accuracy. In general the results obtained were not as high as that currently experienced in the field [5] however the improvement in the base classification accuracy indicates that this could be a viable mechanism for improving classification accuracy. A further analysis will indicate if subsequent improvements can be undertaken which will lead to comparable performance to existing techniques.

It was noted that the 10 fold cross validation did lead to a slightly lower degree of accuracy (88.5%) than utilizing a 4 fold training testing split for our three cluster class (91%). This is most likely attributable to utilizing a 90% training sample producing a small amount of over training. At the same time one of the sub classes in this group achieved a 100% classification accuracy (benign class) while four sub-classes achieved a true positive classification rate in excess of 90% for the sub-classes. The lower classification accuracy in one sub-class potentially indicates that more sub classes exist and that the creation of more clusters and hence cluster targets would improve classification accuracy.

VI. CONCLUSION

A novel technique for classification of abnormalities in digital mammograms has been proposed and investigated in this paper. The proposed technique demonstrates that we have a domain mechanism by which we can create a correlation between our input feature space and our output classification class resulting in an increase in accuracy for the SVM classifier. In this research the improvement was noticeable and the proposed technique achieved 4.5% improvement over standard SVM classifier. The best accuracy was obtained with 3 clusters and 4 fold cross validation. In our future research, more experiments need to be conducted to determine the most suitable kernel function, optimal number of clusters and target classes.

REFERENCES

- [1] A. Yardimci, Soft Computing in Medicine, Applied Soft Computing 9 (3) (2009) 1029-1043.
- [2] H. Cheng, X. Shi, R. Min, L. Ju, X. Cai, H. Du, Approaches for automated detection and classification of masses in mammograms, Pattern Recognition, 39 (4) (2006) 646-668.
- [3] P. Lisboa, A. Taktak, The use of artificial neural networks in decision support in cancer: a systematic review, Neural Networks 19 (2006) 408-415.
- [4] P. Lisboa, A review of evidence of health benefit from artificial neural networks in medical intervention, Neural Networks 15 (2002) 11-39.
- [5] B. Verma, P. McLeod, A. Klevansky, A novel soft cluster neural network for the classification of suspicious areas in digital mammograms, Pattern Recognition 22 (9) (2009) 1845-1852.
- [6] G. Hortobagyi, J. De la Garza Salazar, K. Pritchard, D. Amadori, R. Haidingers, C. Hudis, H. Khaled, M. Liu, M. Martin, M. Namer, J. O'Shaughnessy, Z. Shen, K. Albain, The global breast cancer burden: variations in epidemiology and survival, Clinical Breast Cancer 6 (2005) 391-401.
- [7] R. Nishikawa, M. Kallergi, Computer-aided detection, in its present form, is not an effective aid for screening mammography, Medical Physics 33 (2006) 811-814.
- [8] A. Malich, S. Schmidt, D. Fischer, M. Facius, W. Kaiser, The performance of computer-aided detection when analysing prior mammograms of newly detected breast cancers with special focus on the time interval from initial imaging to detection, European Journal of Radiology 69 (3) (2008) 574-578.
- [9] S. Cohen, N. Intrator, A hybrid projection based on radial basis function architecture: initial values and global optimization, Pattern Analysis and Applications 5 (2) (2002) 113-120.
- [10] M. Heath, K. Bowyer, D. Kopans, R. Moore, P. Kegelmeyer, The digital database for screening mammography, IWDM-2000, Medical Physics Publishing, 2001.
- [11] D. Manrique, J. Rios, A. Rodriguez-Paton, Evolutionary system for automatically constructing and adapting radial basis function networks, Neurocomputing 69 (2006) 2268 – 2283.
- [12] S. Halkiotis, T. Botsis, M. Rangoussi, Automatic detection of clustered microcalcifications in digital mammograms using mathematical morphology and neural networks, Signal Processing 87 (2007) 1559-1568.
- [13] H. Georgiou, M. Mavrofarakis, N. Dimitropoulos, D. Cavouras, S. Theodoridis, Multiscaled morphological features for the characterization of mammographic masses using statistical classification schemes, Artificial Intelligence in Medicine 41 (2007) pp. 39-55.
- [14] C. Varela, P. Tahoces, A. Mendez, M. Souto, J. Vidal, Computerized detection of breast masses in digitized mammograms, Comput Biol Med 37 (2) (2007) 214-216.
- [15] R. Acharya, U. Ng, Y. Chang, J. Yang, G. Kaw, Computer-based identification of breast cancer using digitized mammograms, Journal of Medical Systems 32 (6) (2008) 499-507.
- [16] B. Verma, Novel network architecture and learning algorithm for the classification of mass abnormalities in digitized mammograms, Artificial Intelligence in Medicine 42 (1) (2008) 67-79.
- [17] C. Hsu, C. Chang, C. Lin, A Practical Guide to Support Vector Classification (2009), viewed 16th July 2009 <http://www.csie.ntu.tw/~cjlin>
- [18] I. Christoyianni, A. Kotras, E. Dermatus, G. Kokkinakis, Computer aided diagnosis of breast cancer in digitized mammograms, Computerised Medical Imaging and Graphics 26 (2002) 309-319.
- [19] H. Alto, R. Rangayyan, J. Desautels, Content-based retrieval and analysis of mammographic masses, Journal of Electronic Imaging 14 (023016) (2005) 1-17.
- [20] T. Andre, R. Rangayyan, Classification of breast masses in mammograms using neural networks with shape, edge sharpness, and texture features, Journal of Electronic Imaging 15 (1) (2006), viewed 12 January 2008, doi:10.1117/1.2178271.
- [21] I. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques 2nd Edition, 2005, Morgan Kaufmann.