SVM Approach to Breast Cancer Classification

Mihir Sewak[1], Priyanka Vaidya[1], Chien-Chung Chan[2], Zhong-Hui Duan[2, 3]
[1]Department of Biomedical Engineering, the University of Akron
[2]Department of Computer Science, the University of Akron
[3]Integrated Biosciences Program, the University of Akron
mss41@uakron.edu; psv2@uakron.edu; chan@uakron.edu; duan@uakron.edu

Abstract

The purpose of the proposed study was to provide a solution to the Wisconsin Diagnostic Breast Cancer (WDBC) classification problem. The WDBC dataset, provided by the University of Wisconsin Hospital, was derived from a group of images using Fine Needle Aspiration (biopsies) of the breast. An ensemble of Support Vector Machines (SVM's) was employed in this study. Support vectors with linear, polynomial and RBF kernel functions were trained using a fraction of the WDBC dataset as a training set. The five top performing models were recruited into the ensemble. The classification was then carried out using the majority opinion of the ensemble. The SVM ensemble successfully classified more than 99 percent of the testing data and in the process yielded 100 percent benign tumor prediction accuracy.

Keywords: *Support Vector Machines; Breast Cancer; Machine Learning, Ensemble, Classifiers*

1. Introduction

Breast cancer is the second most leading cause of cancer death in women (after lung cancer). In 2005 breast cancer alone caused 502000 deaths worldwide [1]. It appears in women in the form of lumps or tumors in the breast. Tumors can either be malignant or benign. However, differentiating a malignant tumor from a benign one is a very tedious task due to the structural similarities between the two (Figure 1). It is an extremely critical and time consuming task for the physician to accurately identify the structural differences. Accurate classification is vital as the potency of the cytotoxic drugs administered during treatment can be life threatening. The question remains if there could be an automated technique that could relieve the physician of the tedious task of distinguishing a malignant tumor from a benign one. The following paper is an effort to determine whether this task of automating the prediction process can be achieved with acceptable accuracy.

2. Dataset and Preprocessing

Dataset Details:

In this study we used the dataset provided by researchers at the University of Wisconsin. Dr. Wolberg, at the University Of Wisconsin Hospital, first created the group of images using Fine-Needle Aspiration (FNA) biopsies of the breast. Image processing was then applied on the set of images to come up with the WDBC dataset. The dataset was obtained from the University of California Irvine (UCI) Machine Learning Repository [2].

The features in this dataset were computed from digitized FNA samples. A portion of the well differentiated cell was scanned using a digital camera.

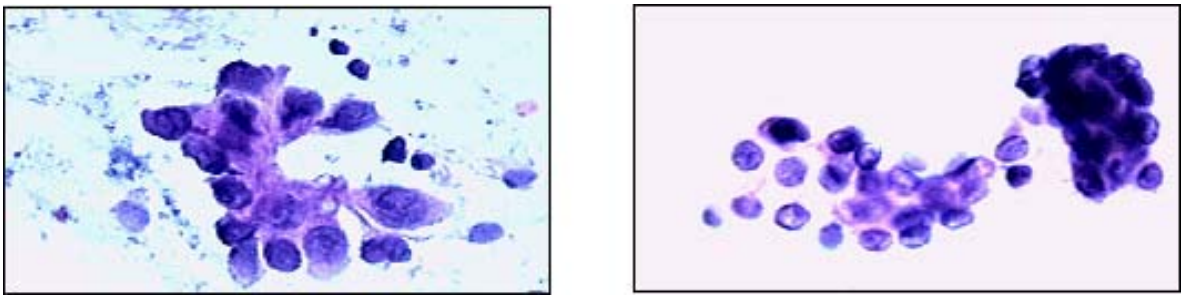

Figure 1: Fine Needle Biopsies of Breast. Malignant (left) and benign (right) breast tumors [4]

0-7695-3039-7/07 $25.00 © 2007 IEEE
DOI 10.1109/IMSCCS.2007.46

The researchers used an image analysis software system "*Xcyt*" to isolate the individual nuclei. An approximate boundary of each nucleus was provided as an input and taken to convergence to the exact nuclear boundary using a semi-automatic process called "*snakes*". In this process of computerized image analysis, the morphometric analysis of cell nuclei to quantify predictive features such as size, shape and texture were carried out [3].

The desired quantification of nuclear shape requires a very precise representation of boundaries. These are generated with the aid of a deformable spline technique known as a '*snake'*. The *snake* seeks to minimize an energy function defined over the arc length of the curve. The energy function is defined in such a way that the minimum value should occur when the curve accurately corresponds to the boundary of a nucleus. The second stage involves the use of these features in inductive machine learning techniques, which use cases with a known (or partially known) outcome to build a mapping from the input features to the decision variable of interest. In order to evaluate the size, shape and texture of each cell nuclei, 10 characteristics were derived namely the radius, perimeter, area, compactness, smoothness, concavity, concave points, symmetry, fractal dimension and texture.

(1) *Radius* was computed by averaging the length of radial line segments, which are lines from the center of mass of the boundary to each of the boundary points.

(2) *Perimeter wa*s measured as the sum of the distances between consecutive boundary points.

(3) *Area* was measured by counting the number of pixels on the interior of the boundary and adding one half of the pixels on to the perimeter to compensate for digitization error.

(4) *Compactness* combined the perimeter and the area to give a measure of the compactness of the cell, calculated as (Perimeter^2) / area

(5) *Smoothness* was quantified by measuring the difference between the length of each radial line and the mean length of the two radial lines surrounding it.

(6) *Concavity* was captured by measuring the size of the indentations in the boundary of the cell nucleus.

(7) *Concave points* were similar to concavity but counted only the number of boundary points lying on the concave regions of the boundary

(8) *Symmetry* was measured by finding the relative difference in length between pairs of line segments perpendicular to the major axis of the contour of the cell nucleus.

(9) *Fractal dimension* was approximated using the "coastline approximation". The perimeter of the nucleus was measured using increasingly large rulers. Plotting the values on a log-log scale and measuring the downward slope gives the negative of an approximation to the fractal dimension.

(10) *Texture* was measured by finding the variance of the gray scale intensities in the component pixels.

The mean value, standard error, and the extreme value of each characteristic were computed for each image, resulting in 30 features of 569 images representing 357 benign and 212 malignant cases. Of the 30 features, the mean and worst case values of Radius, Texture, Parameter, Area, Standard Error of Area, were large-valued.

Data Preprocessing:

To prevent data in the greater numeric range dominating those in the smaller range, we scaled these features in the range 0 to 1 using the standard formula:

$$X_{new} = \frac{(X_{old} - X_{min})}{(X_{max} - X_{min})} \quad (1)$$

Tables 2 and 3 present some data points before and after applying the scaling transformation. The total of all the mean, standard and largest values of the data form a set of 30 attributes and have been plotted from a scale of 0 to 1 in Figure 2, to help visualize the nature of the attribute values for benign and malignant tumors.

Table 1: Unscaled data

Radius	Texture	Parameter	Area
13.54	14.36	87.46	566.3
13.08	15.71	85.63	520
9.504	12.44	60.34	273.9
13.03	18.42	82.61	523.8

Table 2: Values scaled between 0 and 1

Radius	Texture	Parameter	Area
0.310426	0.157254	0.301776	0.179343
0.288655	0.202908	0.28913	0.159703
0.119409	0.092323	0.114367	0.055313
0.286289	0.294555	0.268261	0.161315

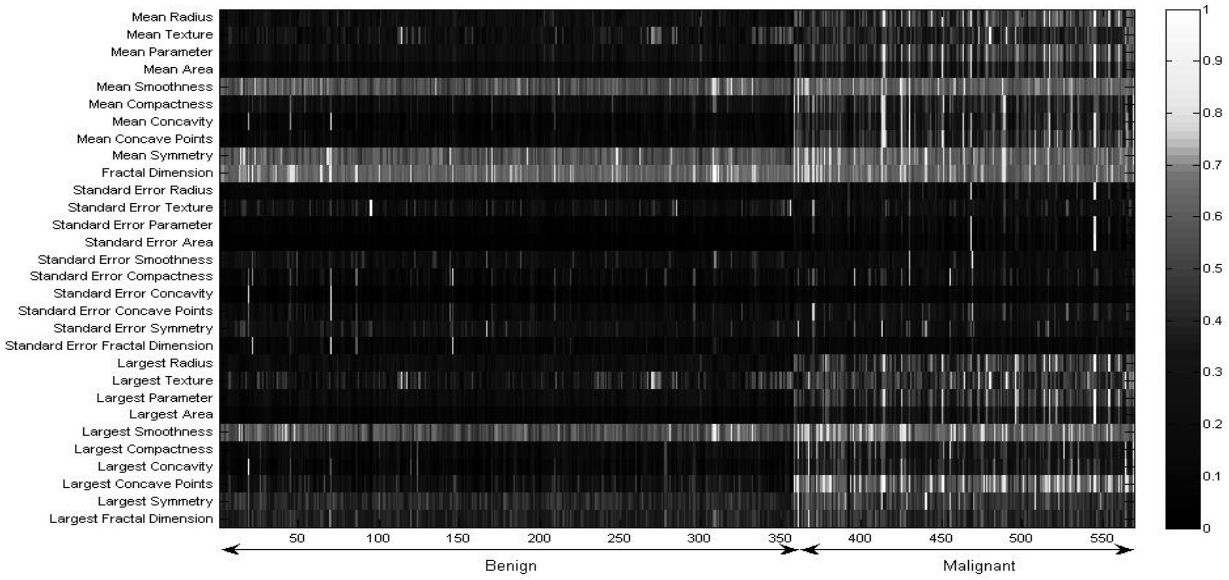

Figure 2: Intensity plot of dataset attributes for both benign and malignant tumors

3. Support Vector Machines

Support Vector Machines (SVM) is a supervised machine-learning tool with a wide application in classification studies. It has been widely used for solving problems in pattern recognition, classification and regression. The SVMs work on an underlying principle, which is to insert a hyper-plane between the classes and orient it in such a way so as to keep it at the maximum distance from the nearest data points as seen in Figure 3. These data points, which appear closest to the hyper-plane, are known as Support Vectors.

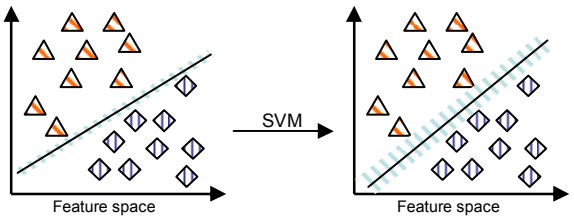

Figure 3: Maximum margin linear classifier

Consider a training sample set with n tuples:

$\{(x_1, y_1),(x_2, y_2),...(x_n, y_n)\}$

Where $x = [x_1, x_2 ... x_n]$ are n data points of the training set, each of which belong to class $y_i \in \{+1, -1\}$

The equation of the hyper-plane can then be represented as:

$$w^T \cdot x + b = 0, \quad (2)$$

where $w = [w_1, w_2 ... w_n]$ is a weight vector and b is a bias. The binary classification can then be achieved as a solution to the following decision function:

$$D(x) = \text{sign}(w^T \cdot x + b) \quad (3)$$

Chen et al. [5] have proven that an optimal hyper-plane (one with maximum margin of separation and without error) is one, which minimizes the cost function:

$$\Phi(w) = \frac{1}{2} w^T \cdot w \quad (4)$$

Subject to the constraint:

$$y_i(w^T \cdot x_i + b) \geq 1, i = 1 : n \quad (5)$$

Because of the convex nature of the cost function, Lagrange multipliers $(\alpha_1, \alpha_2 ... \alpha_n)$ can be used to minimize the problem of constrained optimization, by weighing each data point according to its significance in determining the contrasting information of the two classes [6]. Yao et al. [7] have shown that the optimization problem with the inclusion of the Lagrange multipliers can be rewritten as:

$$\max L(w,b,\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \quad (6)$$

Subject to:

$$\alpha_i \geq 0 \text{ and } \sum_{i=1}^{n} \alpha_i y_i = 0 \quad (7)$$

The decision function with the inclusion of the Lagrange multipliers can then be represented as:

$$D(x) = sign(\sum_{i=1}^{n} \alpha_i y_i (x_i \cdot x) + b) \quad (8)$$

Another important feature of SVMs is that they can be used to solve non-linear classification problems through a kernel function. The kernel function maps two classes of data points in a lower dimensional feature space onto a high dimensional space so that the two sets of data points can be separated using a hyper-plane. In other words, the kernel function converts non-linear classification problems to linear classification problems as depicted in Figure 4.

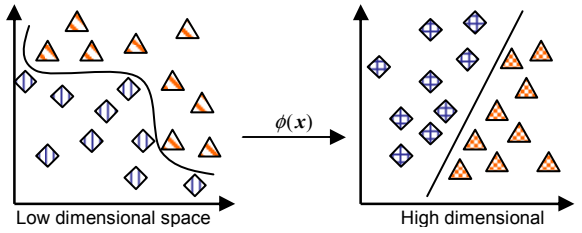

Figure 4: Mapping data points using function ϕ (x)

If $\Phi(x)$ is a transformation function that converts the lower dimension data points to the higher dimensional feature space then the Kernel Function $K(x, y) = \Phi(x)\Phi(y)$ is introduced which performs the necessary transformation [8]. The decision function (8) can then be represented as:

$$D(x) = sign(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b) \quad (9)$$

Many SVM tools were considered for the sake of implementation, some of them being SVMLite, built-in SVM functions from MATLAB and LibSVM etc. However we decided to go ahead with LibSVM because it overcame many limitations of the other tools, in terms of time complexity and ease of use etc. LIBSVM is an open source tool provided by Chih-Chung Chang and Chih-Jen Lin of National Taiwan University [9]. The toolkit comes in various environment dependent versions.

4. Methodology

The preprocessed dataset was subjected to a 10 fold cross validation for purposes of selecting the ensemble of support vectors. The data of 569 samples was divided into 10 groups of roughly 57 samples per group. Nine groups were used in training and the tenth group was used for the prediction. The above process was repeated ten times for the SVM kernel functions (Linear Function, Polynomial Function, and Radial Basis Function). In this way, we obtained 30 models the accuracies of which are documented in Table 3.

Table 3: Prediction accuracy for the 30 trained models. Highlighted models were recruited in the ensemble

Function	Model No.	Linear	Poly.	R.B.F
Prediction Accuracy of Trained Models	1	96.55	100	98.28
	2	94.64	96.43	94.64
	3	93.1	94.83	93.1
	4	98.25	98.25	96.49
	5	96.49	98.24	94.73
	6	96.49	100	96.49
	7	92.86	98.21	94.64
	8	94.74	92.98	92.98
	9	96.49	96.49	94.73
	10	98.21	94.64	98.21
Average Accuracy		95.78	97.01	95.43

The five highlighted models were then recruited into an ensemble. The models were selected on the basis of their prediction accuracies.

The entire dataset of 569 values was passed to the five models and the group decision was considered to decide upon the final output. Figure 5 shows the overall structures of the ensemble formation and usage. Table 4 documents the parameters returned by the LIBSVM tool, details of which can be found on the tool website [9].

Table 4: LIBSVM parameters of the top five models

Model	Support Vectors	Rho
1	43	-4.1839
2	40	-3.8721
3	43	-4.1009
4	102	-6.4274
5	45	-3.6778

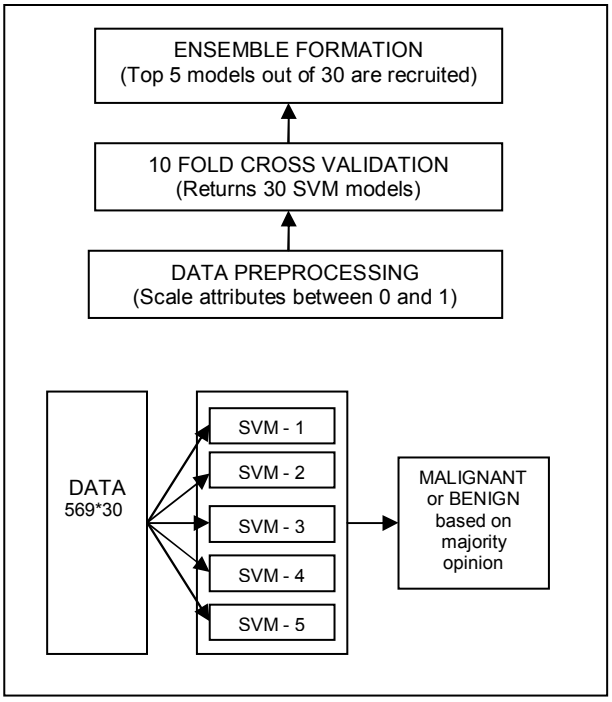

Figure 5: Ensemble Formation and Usage

The models were tested using the complete 569 samples as test data. The results of the ensemble were based on majority voting system.

The system was able to accurately predict 565 of the 569 samples yielding an accuracy of 99.29 percent. The classifier was able to perfectly classify all the benign samples in their proper category, whereas on four out of 212 times it falsely classified malignant cases into the benign category. Table 5 presents the accuracy details of the individual ensemble members along with the group decision and Table 6 present the confusion matrix along with the performance measures of the ensemble of SVMs.

Table 5: Average prediction accuracy compared with the accuracy of the ensemble

MODEL	TRUE	TOTAL	ACC.	AVG. ACC.
1	565	569	99.29	
2	565	569	99.29	
3	564	569	99.12	98.56%
4	546	569	95.96	
5	564	569	99.12	
Ensemble	565	569		99.29%

5. Results and Discussions

For the purpose of the recruitment, a 10-fold Cross-validation approach obtained 30 models, ten each for the linear, polynomial and RBF kernel functions. The top five models, which were inducted into the ensemble, predicted with at least 98 percent accuracy. When the performance of the five models was considered individually, they carried out the prediction with an average accuracy of 98.56.

However when the majority decision was considered, the prediction accuracy obtained was 99.29 percent.

Table 6: Confusion matrix, along with main performance measures for the ensemble

		PREDICTED	
		Benign	Malignant
ACTUAL	Benign	357 (TP)	0 (FN)
	Malignant	4 (FP)	208 (TN)

ACCURACY	$100*\frac{(TP+TN)}{(TP+FP+TN+FN)}$	99.29%
POSITIVE PREDICTIVE VALUE	$100*\frac{(TP)}{(TP+FP)}$	98.89%
NEGATIVE PREDICTIVE VALUE	$100*\frac{(TN)}{(TN+FN)}$	100%
SPECIFICITY	$100*\frac{(TN)}{(TN+FP)}$	98.11%
SENSITIVITY	$100*\frac{(TP)}{(TP+FN)}$	100 %

Two members of the ensemble were able to predict with the same accuracy as that of the recruited group. Thus the ensemble of SVMs performed well to distinguish between benign and malignant characteristics of the WDBC dataset.

Comparing our approach with the other efforts put in towards classifying this data, the results obtained by the proposed study, easily outmatched the other studied approaches.

Anagnostopoulos et al. [10] used a derivative of the probabilistic neural networks (PNN) technique to obtain a classification accuracy of 97.9 percent. Yang et al. [11] employed another derivative of probabilistic neural networks to yield a 98.5 percent classification rate. Li et al. [12] with their technique of employing support vectors for nearest neighbor rules obtained an accuracy of 95.61 percent. Mu et al. [13] applied a Support Vector Machine based classifier using radial basis functions (RBF) networks and self organizing maps (SOMs) for the same dataset obtaining 98.4 percent accuracy. Wolberg et al. [14] made use of inductive machine learning and logistic regression. The logistic regression accuracy of the cross validation was 96.2 percent and the inductive machine learning cross validated classification accuracy was 97.5 percent. Future scope for this work could possibly be a fusion of the SVM classification results of a gene expression pattern in peripheral blood cells with the existing classification. This architecture can then fuse the outputs of different classifiers for more specificity and a more comprehensive output. The successful utilization of this system means that it could be used for diagnostic purposes of breast tumors using Fine-Needle Aspiration data. It will be interesting to see how this system works with the introduction of newer test samples.

6. References

[1] World Health Organization Fact Sheet, (2006), Cancer, http://www.who.int/mediacentre/factsheets/fs297/en/

[2] Asuncion, A & Newman, D.J. (2007). UCI Machine Learning Repository: http://www.ics.uci.edu/~mlearn/MLRepository.html. Irvine, CA: University of California, Department of Information and Computer Science.

[3] Street, W. N. (2000), "Xcyt: A System for Remote Cytological Diagnosis and Prognosis of Breast Cancer", In L.C. Jain, editor, Soft Computing Techniques in Breast Cancer Prognosis and Diagnosis, World Scientific Publishing, Singapore, Pages: 297-322

[4] Street, W., Wolberg, W. and Mangasarian, O. (1994), "Machine Learning Techniques to Diagnose Breast Cancer from Image-Processed Nuclear Features of Fine Needle Aspirates", *Cancer Letters* Vol. 77 pages:163-171

[5] Chen, Y., Wang, G., Dong, S. (2003), "Learning with progressive transductive support vector machine", Pattern Recognition Letters, Vol. 24, Pages: 1845-1855

[6] Brown, M., Gunn, S., R., Lewis, H., G. (1999), "Support vector machines for optimal classification and spectral unmixing", Ecological Modelling, Vol. 120, pages 167-179

[7] Yao, C., Yu, P. (2006), "Fuzzy regression based on asymmetric support vector machines", Applied Mathematics and Computation, Vol. 182 Pages: 175-193

[8] Cheng, J., Yu, D., Yang, Y. (2005), "Application of support vector regression machines to the processing of end effects of Hilbert–Huang transform, Mechanical Systems and Signal Processing", Vol. 21, Issue 3, pages: 1197-1211

[9] Chang, C. and Lin, C. (2001), "LIBSVM: a library for support vector Machines", Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[10] Anagnostopoilosi, I., Anagnostopoilosi, C., Vergados, D., Rouskas, A. and Kormentzas G. (2006), "The Wisconsin Breast Cancer Problem: Diagnosis and TTR/DFS time prognosis using probabilistic and generalized regression information classifiers", *Oncology Reports,* Vol. 15 Pages: 975-981

[11] Yang, Z., Lu, W., Yu, D. and Harrison, R. (2000), "Detecting False Benign in Breast Cancer Diagnosis", *Neural Networks, Proceedings of the IEEE-INNS-ENNS International Joint Conference*, Vol 3, Pages: 655 – 658

[12] Li, Y., Hu, Z., Cai, Y. and Zhang, W. (2005), "Support Vector based Prototype selection Method for Nearest Neighbor Rules", *Lecture Notes in Computer Science 3610,* pages: 528-535

[13] Mu, T., Nandi, A. (2007), "Breast Cancer detection from FNA using SVM with different parameter tuning systems and SOM-RBF classifier", *Journal of the Franklin Institute*, Vol. 344, pages: 285-311

[14] Wolberg, W.,H., Street, W.,N., Mangasarian O., L. (1993), "Image Analysis and machine learning applied to breast cancer diagnosis and prognosis", *Anal., Quant., Cytol., Histol,* Vol. 15 pages: 396-404