A Support Vector Machine Approach to Breast Cancer Diagnosis and Prognosis

Elias Zafiropoulos, Ilias Maglogiannis, Ioannis Anagnostopoulos [1]
1 Department of Information and Communication Systems Engineering,
University of the Aegean,
GR 83200 Karlovasi, Samos, Greece

Abstract. In recent years, computational diagnostic tools and artificial intelligence techniques provide automated procedures for objective judgments by making use of quantitative measures and machine learning. The paper presents a Support Vector Machine (SVM) approach for the prognosis and diagnosis of breast cancer implemented on the Wisconsin Diagnostic Breast Cancer (WDBC) and the Wisconsin Prognostic Breast Cancer (WPBC) datasets found in literature. The SVM algorithm performs excellently in both problems for the case study datasets, exhibiting high accuracy, sensitivity and specificity indices.

1 Introduction

The implementation of training algorithms in the prognosis and diagnosis of cancer is a research area of great interest, while significant research work has been published in literature, basically in the area of neural networks [1], [2], [3], [4]. In the present re-search work, a SVM model is implemented for the breast cancer diagnosis and prognosis problem using the Wisconsin Diagnostic Breast Cancer (WDBC) as well as the Wisconsin Prognostic Breast Cancer (WPBC) datasets, which are publicly available at http://ftp.ics.uci.edu/pub/machine-learning-databases/breast-cancer-wisconsin/. These datasets involve measurements taken according the Fine Needle Aspirate (FNA) test. The role of diagnosis is to provide a distinction between the malignant and benign breast masses. In case that a patient is diagnosed with breast cancer, the malignant mass must be excised. After this or a different post-operative procedure, a prediction of the expected course of the disease must be determined. However, prognostic pre-diction does not belong either on the classic learning paradigms of function approximation or classification. This is due to a patient can be classified as a "recur" case (instance) if the disease is observed, while there is not a threshold point at which the patient can be considered as a "non-recur" case. The data are therefore censored since a time to recur for only a subset of

Please use the following format when citing this chapter:
Zafiropoulos, Elias, Maglogiannis, Ilias, Anagnostopoulos, Ioannis, 2006, in IFIP International Federation for Information Processing, Volume 204, Artificial Intelligence Applications and Innovations, eds. Maglogiannis, I., Karpouzis, K., Bramer, M., (Boston: Springer), pp. 500–507

patients is known. For the other patients, the length of time after treatment during which malignant masses are not found is known. This time interval is the disease free survival (DFS) time, which can be reported for an individual patient or for a study population. In particular, the right endpoints of the recurrence time intervals are right censored, as some patients will inevitably change hospital, doctors or die of other unrelated with the cancer causes. The prognosis of the specific time interval is considered a difficult problem since the training data are right censored [1], [2], [3], [4].

In the present paper, the Support Vector Machines (SVM) algorithm is implemented for the breast cancer diagnosis and prognosis problem and the WPBC – WDBC data are used as a case study [5], [6], [7], [8]. The SVM algorithm performed excellently in both prognosis and diagnosis problems for the WPBC/WDBC datasets exhibiting high accuracy, sensitivity and specificity indices. In Section 2, all the details concerning the medical data characteristics and the problem formulation for each dataset in the cases of prognosis and diagnosis are presented. Section 3 contains the basic principles of the SVM algorithm for data classification. Section 4 presents the proposed approach for prognosis and diagnosis of the case study datasets and all the corresponding results, while in Section 5 the paper is concluded.

2. Medical Data Characteristics and Problem Formulation

The WDBC and WPBC datasets are the results of the efforts made at the University of Wisconsin Hospital for the diagnosis and prognosis of breast tumours solely based on FNA test. This test involves fluid extraction from a breast mass using a small-gauge needle and then visual inspection of the fluid under a microscope. Figure 1 depicts two images, which were taken from fine needle biopsies of breast as appeared in [9].

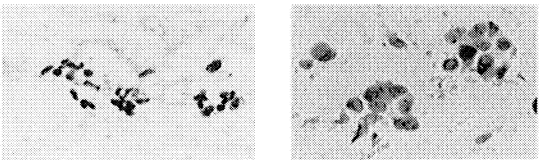

Fig. 1. Images taken using the FNA test: (a) Benign, (b) Malignant

The WDBC dataset consist of 569 instances (357 benign – 212 malignant), where each one represents FNA test measurements for one diagnosis case. For this dataset each instance has 32 attributes, where the first two attributes correspond to a unique identification number and the diagnosis status (benign / malignant). The rest 30 features are computations for ten real-valued features, along with their mean, standard error and the mean of the three largest values ("worst" value) for each cell nucleus respectively. These ten real values, which are depicted at Table 1, are computed from a digitized image of a fine needle aspirate (FNA) of breast tumour, describing characteristics of the cell nuclei present in the image and are recorded with four significant digits.

The WPBC dataset consists of 198 instances (151 non-recur - 47 recur), where each one represents follow-up data for one breast cancer case. Each instance has 35 attributes, where the first three attributes correspond to a unique identification number and to the prognosis status (recur / non-recur) following by the recurrence time (Time to Recur - TTR) or the DFS time respectively. Then they follow the above-mentioned 30 features, while the last two attributes are the diameter of the excised tumour (in cm) and the number of positive axillary lymph nodes observed at time of surgery. Four instances were not included in the training/testing set since the Lymph node values were missing. For the addressed problem, both WDBC and WPBC datasets were used in several publications in the medical literature [10], [11], [12], [13]. In addition, due to their consistency and robust creation, these datasets are also used for verification purposes over the classification or prediction performance of information systems in other scientific areas [14], [15].

Table 1. WDBC/WPBC cell nuclei characteristics attributes.

Cell Nuclei Characteristics
1. radius [mean of distances from centre to points on the perimeter],
2. texture [standard deviation of grey-scale values],
3. perimeter,
4. area,
5. smoothness [local variation in radius lengths],
6. compactness [((perimeter)2 / area) - 1],
7. concavity [severity of concave portions of the contour],
8. concave points [number of concave portions of the contour],
9. symmetry,
10. fractal dimension ["coastline approximation" – 1]

3 Principles of the Support Vector Machines Algorithm for Data Classification

The Support Vector Machines (SVMs) is a novel algorithm for data classification and regression which allows the expansion of the information provided by a training dataset as a linear combination of a subset of the data in the training set (support vectors) [5], [6]. These vectors locate a hypersurface that separates the input data with a very good degree of generalization. The SVM algorithm is a learning machine; there-fore it is based on training, testing and performance evaluation, which are common steps in every learning procedure. Training involves optimization of a convex cost function where there are no local minima to complicate the learning process. Testing is based on the model evaluation using the support vectors to classify a test dataset. Performance is based on error rate determination as test dataset size tends to infinity.

The mathematical formulation of the Support Vector Machine algorithm for data classification and regression is presented extensively in literature [5], [6], [7]. A critical issue is the selection of a suitable kernel function that will transform the

initially non-separable data in a new feature space where they are separable. Several kernel functions can be used, such as the following:

$$k(\mathbf{X}_i,\mathbf{X}_j)=e^{-\|\mathbf{X}_i-\mathbf{X}_j\|^2/2\sigma^2} \quad \text{(Gaussian RBF kernel)} \quad (1)$$

$$k(\mathbf{X}_i,\mathbf{X}_j)=(\mathbf{X}_i\cdot\mathbf{X}_j+m)^p \quad \text{(polynomial kernel)} \quad (2)$$

4 The proposed approach of the SVM Algorithm Classification Problem for Prognosis and Diagnosis of Breast Cancer

4.1 Classification of the WPBC patient data based on the Disease Free or Recurrence Time (prognosis)

The SVM algorithm has been implemented in the case of the WPBC instances for data classification according to the recurrence or the DFS time. The WPBC instances were divided over four classes, namely C1, C2, C3 and C4, according to the value of the recurrence or the DFS time. In other words, C1 corresponds to the instances, in which the DFS time or the recurrence time was between 1 and 12 months, while C2, C3 and C4 correspond to intervals between 1-3 years, 3-6 years and more than 6 years. Table 2 depicts the amount of the WPBC dataset instances in respect to the above-mentioned categorization. The first column indicates the time interval class, while the second and the third columns present the amount of instances, when the tumour recurred (NR) and the amount of instances when the tumour did not recur (NN).

Table 2. WPBC instances according to the categorized interval time and prognosis status

Class	Interval time	N_R	N_N	Total
C_1	Less than 1 year	20	23	43
C_2	1 year – 3 years	14	34	48
C_3	3 years – 6 years	7	48	55
C_4	More than 6 years	5	43	48
			Total	194

Based on the categorization of the WPBC instances in the four intervals depicted in Table 2, the SVM algorithm has been applied for the corresponding two-class classification problem of each time interval. The training set and test set originated from the WPBC instances, while the attributes used were the ones depicted in Table 1, together with the "tumour size" and "lymph node status" features found only in the WPBC dataset. Several kernel functions were tried in order to find the least complex function that results in low number of support vectors comparing to the training set and exhibits satisfactory performance in data classification. The top results are depicted in Table 3. The Gaussian Radial Base function (RBF) with sigma=1 exhibits the best performance with accuracy varying from 96.91% to 94.84% for the four time intervals. The SVM algorithm was implemented in Matlab using a Pentium PC at 2.6GHz with 512 MB RAM. The execution time for calculating the support vectors using the Gaussian RBF with sigma=1 as a kernel

function was approximately 12 seconds, while for the rest of the cases with different kernel functions varied for 11 to 14 seconds.

Apart from the accuracy indices presented in Table 3, the performance of a binary classifier can be further evaluated using the sensitivity and specificity indices. Assuming that if an instance belongs to the time interval it is classified positive, otherwise it is classified negative, the sensitivity and specificity indices can be defined as follows:

$$\text{Sensitivity}=TP/(TP+FN) \quad (3)$$

$$\text{Specificity}=TN/(TN+FP) \quad (4)$$

where:

TP (TN) =Number of True Positive (True Negative) classified instances; i.e. instances that the learning machine classifies correctly

FP (FN) =Number of False Positive (False Negative) classified instances; i.e. the learning machine labels the instance as positive (negative) while it is negative (positive)

Table 3: Results of the SVM algorithm using alternative kernel functions for the classification of WPBC cases in each time interval

		Kernel functions				
		Polynomial		**Gaussian RBF**		
		p=2	**p=3**	**σ=1**	**σ=2**	**σ=3**
C1	**No of SVs**	74	91	*104*	93	104
	Errors	152	42	*10*	27	39
	Accuracy (%)	21.65	78.35	*94.84*	86.08	79.90
C2	**No of SVs**	77	106	*122*	103	112
	Errors	46	46	*9*	20	31
	Accuracy (%)	76.29	76.29	*95.36*	89.69	84.02
C3	**No of SVs**	192	97	*108*	99	109
	Errors	55	55	*6*	31	43
	Accuracy (%)	71.65	71.65	*96.91*	84.02	77.84
C4	**No of SVs**	177	78	*100*	93	95
	Errors	143	51	*9*	19	33
	Accuracy (%)	26.29	73.71	*95.36*	90.21	82.99

These indices have been calculated for the SVM learning machine using the Gaussian RBF with σ=1 (sigma=1) and the corresponding results are presented in Table 4. The presented performance indices have been also calculated for the various kernel functions examined in Table 3 and the corresponding results for the kernel functions with the top performance are presented in Figures 3 and 4. In these figures, the specificity and sensitivity indices for the Gaussian RBF (sigma=1) are higher comparing to other kernel functions in all cases of classification in the four time intervals. This fact together with the high total accuracy depicted in Table 3 indicate the proposed SVM learning algorithm with the Gaussian RBF (sigma =1) kernel function as a superior binary classifier of the WPBC instances in the selected time intervals of recurrence or the DFS time.

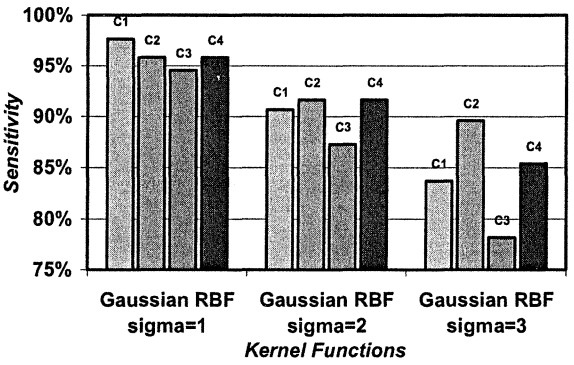

Fig. 3: Sensitivity indices for the SVM learning machine using alternative kernel functions for the classification according to the recurrence or the disease-free time survival (DFS) time.

Table 4: Sensitivity and specificity indexes for the SVM algorithm implementation using the Gaussian radial base function with σ=1 as kernel function

	Time Intervals			
	C1	**C2**	**C3**	**C4**
Positive	43	48	55	48
Negative	151	146	139	146
True Positive Classified	42	46	52	46
True Negative Classified	142	139	136	139
False Positive Classified	9	7	3	7
False Negative Classified	1	2	3	2
SENSITIVITY (%)	97.67	95.83	94.55	95.83
SPECIFICITY (%)	94.04	95.21	97.84	95.21

4.2. Automated Diagnosis of Breast Cancer Based on the WDBC Patient Data

Furthermore, the SVM algorithm has been implemented for the successful automated diagnosis of benign vs. malignant melanoma instances in the case of the WDBC patient data. The training set was constructed by randomly selecting 350 cases out of the WDBC instances, while the complete dataset was used for test set. In this way, the efficiency of the SVM algorithm has been examined using data that have not been used in the train set. Several kernel functions were tried in order to find the least complex function that results in low number of support vectors comparing to the training set and exhibit satisfactory performance in data classification. The best results are presented in Table 5. The Gaussian RBF with σ=0.6 exhibits the best performance with accuracy approximately 90%. Accordingly, the SVM algorithm for the WDBC case was also implemented in Matlab using a Pentium PC at 2.6GHz with 512 MB RAM. The execution time for calculating the support vectors based on the kernel functions of varied from 70 to 74 seconds.

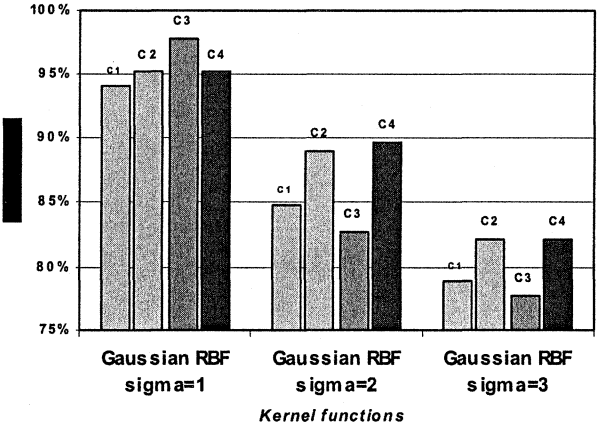

Fig 4: Specificity indices for the SVM learning machine using alternative kernel functions for the classification according to the recurrence or the disease-free time survival (DFS) time.

Table 5: Results of the SVM algorithm for the diagnosis of benign / malignant melanoma for the WDBC cases

Kernel functions	No of SVs	Errors	Accuracy (%)
Gaussian RBF (σ=3)	45	112	80.32%
Gaussian RBF (σ=2)	42	109	80.84%
Gaussian RBF (σ=1)	48	83	85.41%
Gaussian RBF (σ=0.8)	54	73	87.17%
Gaussian RBF (σ=0.6)	*67*	*61*	*89.28%*

The sensitivity and specificity indices, as they have been defined in Eq 12 and 13, have been calculated for all kernel functions presented in Table 5 and the corresponding results are presented in Figure 5, together with the accuracy of each kernel function.

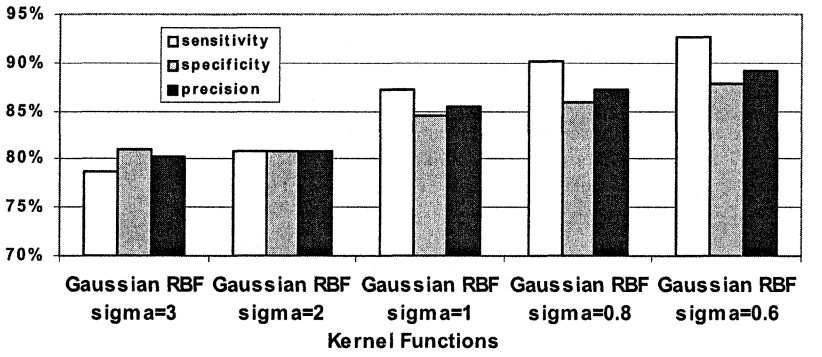

Fig. 5: Sensitivity, specificity and accuracy for the top performance kernel functions in the SVM learning machine for diagnosis of the WDBC instances

5. Conclusions

This paper focuses on the implementation of the SVM algorithm for the diagnosis and prognosis of breast cancer. Firstly, the methodology was implemented for the prognosis problem based on Wisconsin Prognostic Breast Cancer datasets. The SVM algorithm performed excellently, exhibiting high values of accuracy (96.91%), specificity and sensitivity indices. Similarly, automated diagnosis using SVM was implemented for the Wisconsin Diagnostic Breast Cancer datasets and the accuracy was approximately 90%, while sensitivity and specificity indices were also satisfactory.

References

1. Burke H. B., Goodman P.H., et al, Artificial neural networks improve the accuracy of cancer survival prediction, Cancer, Vol. 79, pp. 857-862, 1997.
2. Choong P.L, deSilva C.J.S et al., Entropy maximization networks, An application to breast cancer prognosis, IEEE Transactions on Neural Networks, 1996, 7(3):568-577.
3. Mangasarian et al, "Breast cancer diagnosis and prognosis via linear programming", Operations Research, 43(4), pp. 570-577, July-August 1995.
4. Street W. N., "A neural network model for prognostic prediction", Proceedings of theFifteenth International Conference on Machine Learning, Madison, Wisconsin, Morgan Kaufmann, 1998.
5. Burges C.: A tutorial on support vector machines for pattern recognition [http://www.kernel-machines.org/].
6. Schölkopf B.: Statistical learning and kernel methods [http://research.Microsoft.com/~bsc].
7. Campbell C.: Kernel methods: a survey of current techniques, [http://www.kernel-machines.org/].
8. Maglogiannis I. G., Zafiropoulos E. P. Characterization of digital medical images utilizing support vector machines. BMC Medical Informatics and Decision Making 2004; 4:4.
9. Wolberg W.H., Street W.N., Heisey D.M., and Mangasarian O.L., Computer-derived nuclear features distinguish malignant from benign breast cytology, Human Pathology, 26:792--796, 1995.
10. Tourassi G.D., Markey M.K., Lo J.Y., Floyd Jr. C.E., A neural network approach to breast cancer diagnosis as a constraint satisfaction problem, Med. Phys. Vol.28, pp. 804–811, 2001.
11. Wolberg W.H., Street W.N., Heisey D.M., and Mangasarian O.L., Computer-derived nuclear features distinguish malignant from benign breast cytology, Human Pathology, 26:792--796, 1995.
12. Wolberg W.H., Street W.N., and Mangasarian O.L., Machine learning techniques to diagnose breast cancer from fine-needle aspirates, Cancer Letters 77 (1994) 163-171.
13. Wolberg W.H., Street W.N., and Mangasarian O.L., Image analysis and machine learning applied to breast cancer diagnosis and prognosis, Analytical and Quantitative Cytology and Histology, Vol. 17, No. 2, pages 77-87, April 1995.
14. Hoya T. and Chambers J. A., "Heuristic pattern correction scheme using adaptively trained generalized regression neural networks", IEEE Trans. Neural Networks, vol.12, no.1, pp. 91-100, 2001.
15. Kaban A., Girolami M., Initialized and guided EM-clustering of sparse binary data with application to text based documents, 15th International Conference on Pattern Recognition, Vol.2 pp.744-747, Sept. 2000.