Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 36 (2009) 3240–3247
Expert Systems with Applications
www.elsevier.com/locate/eswa

Support vector machines combined with feature selection for breast cancer diagnosis

Mehmet Fatih Akay *

Department of Electrical and Electronics Engineering, Cukurova University, Adana 01330, Turkey

Abstract

Breast cancer is the second largest cause of cancer deaths among women. At the same time, it is also among the most curable cancer types if it can be diagnosed early. Research efforts have reported with increasing confirmation that the support vector machines (SVM) have greater accurate diagnosis ability. In this paper, breast cancer diagnosis based on a SVM-based method combined with feature selection has been proposed. Experiments have been conducted on different training-test partitions of the Wisconsin breast cancer dataset (WBCD), which is commonly used among researchers who use machine learning methods for breast cancer diagnosis. The performance of the method is evaluated using classification accuracy, sensitivity, specificity, positive and negative predictive values, receiver operating characteristic (ROC) curves and confusion matrix. The results show that the highest classification accuracy (99.51%) is obtained for the SVM model that contains five features, and this is very promising compared to the previously reported results.

Keywords: Breast cancer diagnosis; Support vector machines; Feature selection

1. Introduction

Cancer is a group of diseases in which cells in the body grow, change, and multiply out of control. Usually, cancer is named after the body part in which it originated; thus, breast cancer refers to the erratic growth of cells that originate in the breast tissue. A group of rapidly dividing cells may form a lump or mass of extra tissue. These masses are called tumors. Tumors can either be cancerous (malignant) or non-cancerous (benign). Malignant tumors penetrate and destroy healthy body tissues.

The term, breast cancer, refers to a malignant tumor that has developed from cells in the breast. Breast cancer is the leading cause of death among women between 40 and 55 years of age and is the second overall cause of death among women (exceeded only by lung cancer) (http://www.imaginis.com/breasthealth/breast_cancer.asp, Last Accessed August 2007). According to the World Health Organization, more than 1.2 million women will be diagnosed with breast cancer each year worldwide. Fortunately, the mortality rate from breast cancer has decreased in recent years with an increased emphasis on diagnostic techniques and more effective treatments. A key factor in this trend is the early detection and accurate diagnosis of this disease (West, Mangiameli, Rampal, & West, 2005).

The use of classifier systems in medical diagnosis is increasing gradually. There is no doubt that evaluation of data taken from patients and decisions of experts are the most important factors in diagnosis. However, expert systems and different artificial intelligence techniques for classification also help experts in a great deal. Classification systems can help minimizing possible errors that can be done because of inexperienced experts, and also provide medical data to be examined in shorter time and more detailed.

SVM have been proposed as an effective statistical learning method for classification (Vapnik, 1989). They rely on so called support vectors (SV) to identify the decision

* Corresponding author.
E-mail address: mfakay@cu.edu.tr

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.01.009

boundaries between different classes. SVM are based on a linear machine in a high dimensional feature space, non-linearly related to the input space, which has allowed the development of somewhat fast training techniques, even with a large number of input variables and big training sets. SVM have been used successfully for the solution of many problems including handwritten digit recognition (Scholkopf et al., 1997), object recognition (Pontil & Verri, 1998), speaker identification (Wan & Campbell, 2000), face detection in images (Osuna, Freund, & Girosi, 1997), and text categorization (Joachims, 1999).

When using SVM, three problems are confronted: how to choose the kernel function and optimal input feature subset for SVM, and how to set the best kernel parameters. These problems are crucial because the feature subset choice influences the appropriate kernel parameters and vice versa (Frohlich et al., 2003). Feature selection is an important issue in building classification systems. It is advantageous to limit the number of input features in a classifier to in order to have a good predictive and less computationally intensive model (Zhang, 2000). With a small feature set, the explanation of rationale for the classification decision can be more readily realized.

In this study, SVM with feature selection was used to diagnose the breast cancer. WBCD taken from the University of California at Irvine (UCI) machine learning repository was used for training and testing experiments (ftp://ftp.ics.uci.edu/pub/machine-learning-databases/breast-cancer-wisconsin, Last Accessed August 2007). It was observed that the proposed method yielded the highest classification accuracies (98.53%, 99.02%, and 99.51% for 50–50% of training-test partition, 70–30% of training-test partition, and 80–20% of training-test partition, respectively) for a subset that contained five features. Also, other measures such as the confusion matrix, sensitivity, specificity, positive predictive value, negative predictive value and ROC curves were used to show the performance of SVM with feature selection.

The rest of the paper is organized as follows. Section 2 summarizes the methods and results of previous research on breast cancer diagnosis. Section 3 reviews basic SVM concepts. Section 4 describes the proposed method. Section 5 presents experimental results from using the proposed method to diagnose breast cancer. Finally, Section 6 concludes the paper along with outlining future directions.

2. Related work on breast cancer diagnosis

There has been a lot of research on medical diagnosis of breast cancer with WBCD in literature, and most of them reported high classification accuracies. In Albrecht, Lappas, Vinterbo, Wong, and Ohno-Machado (2002), a learning algorithm that combined logarithmic simulated annealing with the perceptron algorithm was used and the reported accuracy was 98.8%. In Pena-Reyes and Sipper (1999), the classification technique used fuzzy-GA method reaching a classification accuracy of 97.36%. In Setiono (2000), the classification was based on a feed forward neural network rule extraction algorithm. The reported accuracy was 98.10%. (Quinlan, 1996) reached 94.74% classification accuracy using 10-fold cross-validation with C4.5 decision tree method. (Hamiton, Shan, & Cercone, 1996) obtained 94.99% accuracy with RIAC method, while (Ster & Dobnikar, 1996) obtained 96.8% with linear discreet analysis method. The accuracy obtained by Nauck and Kruse (1999) was 95.06% with neuron-fuzzy techniques. In Goodman, Boggess, and Watkins (2002), three different methods, optimized learning vector quantization (LVQ), big LVQ, and artificial immune recognition system (AIRS), were applied and the obtained accuracies were 96.7%, 96.8%, and 97.2%, respectively. In Abonyi and Szeifert (2003), an accuracy of 95.57% was obtained with the application of supervised fuzzy clustering technique. In Polat and Gunes (2007), least square SVM was used and an accuracy of 98.53% was obtained.

3. Support vector machines

3.1. Linear SVM

Consider the problem of separating the set of training vectors belonging to two linearly separable classes,

$$(\mathbf{x}_i, y_i), \quad \mathbf{x}_i \in R^n, \quad y_i \in \{+1, -1\}, \quad i = 1, \ldots, n, \tag{1}$$

where $\mathbf{x}_i$ is a real-valued n-dimensional input vector and y_i is a label that determines the class of $\mathbf{x}_i$. A separating hyperplane is determined by an orthogonal vector $\mathbf{w}$ and a bias b, which identifies the points that satisfy

$$\mathbf{w}.\mathbf{x} + b = 0. \tag{2}$$

The parameters $\mathbf{w}$ and b are constrained by

$$\min_i |\mathbf{w}.\mathbf{x}_i + b| \geqslant 1. \tag{3}$$

A separating hyperplane in canonical form must satisfy the following constraints,

$$y_i(\mathbf{w}.\mathbf{x}_i + b) \geqslant 1, \quad i = 1, 2, \ldots, n. \tag{4}$$

The hyperplane that optimally separates the data is the one that minimizes

$$\Phi(\mathbf{w}) = \frac{1}{2}(\mathbf{w}.\mathbf{w}). \tag{5}$$

Relaxing the constraints of (4) by introducing slack variables $\xi_i \geqslant 0$, $i = 1, 2, \ldots, n$, (4) becomes

$$y_i.(\mathbf{w}.\mathbf{x}_i + b) \geqslant 1 - \xi_i, \quad i = 1, 2, \ldots, n. \tag{6}$$

In this case, the optimization problem becomes

$$\Phi(\mathbf{w}, \xi) = \frac{1}{2}(\mathbf{w}.\mathbf{w}) + C\sum_{i=1}^{n} \xi_i \tag{7}$$

with a user defined positive finite constant C. The solution to the optimization problem in (7), under the constraints of

(6), could be obtained in the saddle point of Lagrangian function

$$L(\mathbf{w},b,\alpha,\xi,\gamma)=\frac{1}{2}(\mathbf{w}.\mathbf{w})+C\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i|y_i(\mathbf{w}.\mathbf{x}_i+b)-1+\xi_i|-\sum_{i=1}^{n}\gamma_i\xi_i, \tag{8}$$

where $\alpha_i \geqslant 0$, $\xi_i \geqslant 0$, $i=1,2,\ldots,n$ are the Lagrange multipliers. The Lagrangian function has to be minimized with respect to $\mathbf{w}$, b, and ξ_i. Classical Lagrangian duality enables the primal problem, (8), to be transformed into its dual problem, which is easier to solve. The dual problem is given by

$$\max_{\alpha}\left[\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j(\mathbf{x}_i.\mathbf{x}_j)\right] \tag{9}$$

with constraints

$$\sum_{i=1}^{n}\alpha_i y_i=0,\quad 0\leqslant\alpha_i\leqslant C,\ \ i=1,2,\ldots,n. \tag{10}$$

This is a classic quadratic optimization problem, for which there exists a unique solution. According to the Kuhn–Tucker theorem of optimization theory (Bertsekas, 1995), the optimal solution satisfies

$$\alpha_i[y_i(\mathbf{w}.\mathbf{x}_i+b)-1]=0,\quad i=1,2,\ldots,n. \tag{11}$$

(11) has non-zero Lagrange multipliers if and only if the points $\mathbf{x}_i$ satisfy

$$y_i(\mathbf{w}.\mathbf{x}_i+b)=1. \tag{12}$$

These points are termed SV. The hyperplane is determined by the SV, which is a small subset of the training vectors. Hence if α_i^* is the non-zero optimal solution, the classifier function can be expressed as

$$f(\mathbf{x})=\operatorname{sgn}\left\{\sum_{i=1}^{n}\alpha_i^* y_i(\mathbf{x}_i.\mathbf{x})+b^*\right\}, \tag{13}$$

where b^* is the solution of (11) for any non-zero α_i^*.

3.2. Non-linear SVM

When a linear boundary is inappropriate SVM can map the input vector into a high dimensional feature space. By defining a non-linear mapping, the SVM construct an optimal separating hyperplane in this higher dimensional space. Usually non-linear mapping is defined as

$$\varphi(\cdot):R^n\rightarrow R^{nh}. \tag{14}$$

In this case, optimal function (9) becomes (15) with the same constraints

$$\max_{\alpha}\left[\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j)\right], \tag{15}$$

where

$$K(\mathbf{x}_i,\mathbf{x}_j)=\{\varphi(\mathbf{x}_i).\varphi(\mathbf{x}_j)\} \tag{16}$$

is the kernel function performing the non-linear mapping into feature space. The kernel function may be any of the symmetric functions that satisfy the Mercel conditions (Courant & Hilbert, 1953). The most commonly used functions are the Radial Basis Function (RBF)

$$K(\mathbf{x}_i,\mathbf{x}_j)=\exp\{-\gamma\mid\mathbf{x}_i-\mathbf{x}_j|^2\} \tag{17}$$

and the Polynomial Function

$$K(\mathbf{x}_i,\mathbf{x}_j)=(\mathbf{x}_i.\mathbf{x}_j+1)^q,\quad q=1,2,\ldots, \tag{18}$$

where the parameters γ and q in (17) and (18), respectively must be preset.

4. Methodology and experiments

4.1. Breast cancer dataset

We have used WBCD taken from the UCI machine learning repository in our experiments. This dataset is commonly used among researchers who use machine learning methods for breast cancer classification, so it provides us to compare the performance of our method with that of others. The dataset contains 683 samples taken from needle aspirates from human breast cancer tissue. It consists of nine features, each of which is represented as an integer between 1 and 10. The features are; clump thickness (F_1), uniformity of cell size (F_2), uniformity of cell shape (F_3), marginal adhesion (F_4), single epithelial cell size (F_5), bare nucleoi (F_6), bland chromatin (F_7), normal nuclei (F_8), and mitoses (F_9). Four hundred and forty four samples of the dataset belong to benign class, and the rest are of malignant class.

4.2. Feature Selection

Feature selection is an important issue in building classification systems. It is advantageous to limit the number of input features in a classifier in order to have a good predictive and less computationally intensive model. In the area of medical diagnosis, a small feature subset means lower test and diagnostic costs. F-score (Chen & Lin, 2005) is a simple technique which measures the discrimination of two sets of real numbers. Given training vectors x_k, $k=1,2,\ldots,m$, if the number of positive and negative instances are n_+ and n_-, respectively, then the F-score of the ith feature is defined as

$$F_i=\frac{(\bar{x}_i^{(+)}-\bar{x}_i)^2+(\bar{x}_i^{(-)}-\bar{x}_i)^2}{\frac{1}{n_+-1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)}-\bar{x}_i^{(+)})^2+\frac{1}{n_--1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)}-\bar{x}_i^{(-)})^2}, \tag{19}$$

where $\bar{x}_i,\bar{x}_i^{(+)},\bar{x}_i^{(-)}$ are the average of the i^{th} feature of the whole, positive, and negative datasets, respectively; $x_{k,i}^{(+)}$ is the ith feature of the kth positive instance, and $x_{k,i}^{(-)}$ is the

*i*th feature of the *k*th negative instance. The numerator indicates the discrimination between the positive and negative sets, and the denominator indicates the one within each of the two sets. The larger the *F*-score is, the more likely this feature is more discriminative.

4.3. Setting model parameters

In addition to the feature subset selection, kernel selection and setting appropriate kernel parameters can greatly improve the SVM classification accuracy. In this work, we chose the RBF kernel defined by (17). The parameters that should be optimized for the RBF kernel are the penalty parameter C and the kernel function parameter γ.

For median-sized problems, the grid search approach is an efficient way to find the best C and γ (Hsu, Chang, & Lin, 2003). In grid search, pairs of (C, γ) are tried and the one with the best cross-validation accuracy is chosen. To improve the generalization ability, grid search uses a cross-validation process. That is, for each k subsets of the dataset D, create a training set $T = D - k$, then run a cross-validation (CV) process as follows:

1. Consider a grid space of (C,γ) with $\log_2 C \in \{-5,-4,\ldots,15\}$ and $\log_2\gamma \in \{-15,-11,\ldots,3\}$.
2. For each pair (C,γ) in the search space, conduct k-fold CV on the training set.
3. Choose the parameter (C,γ) that leads to the highest overall CV classification rate.
4. Use the best parameter to create a model for training the dataset.

4.4. SVM model with grid search and feature selection

Fig. 1 shows the block diagram of our SVM-based model using grid search to find the optimized model parameters and *F*-score calculation to select input features. Initially, the *F*-score of each feature in the training set is calculated and these scores are sorted in descending order. In the second phase, a subset of the original training set is generated by including the features with top N *F*-scores, where $N = 1,2,\ldots,m$ and m is the total number of features. Once the subset is generated, a grid search using 10-fold CV is carried out in the third phase to find the optimized values of (C,γ). In the fourth stage, the subset is trained with the values of (C,γ) found in the previous stage and the SVM predictor model is obtained. In the last stage, this model is used to predict labels in the test subset. This procedure is carried out until all the features appear in the subset according to their *F*-scores.

4.5. Measures for performance evaluation

We have used several measures in order to evaluate the effectiveness of our method. These measures are classification accuracy, sensitivity, specificity, positive predictive value, negative predictive value, ROC curves and confusion matrix. A confusion matrix (Kohavi & Provost, 1998) contains information about actual and predicted classifications done by a classification system. Table 1 shows the confusion matrix for a two class classifier. Classification accuracy, sensitivity, specificity, positive predictive value and negative predictive value can be defined by using the elements of the confusion matrix as

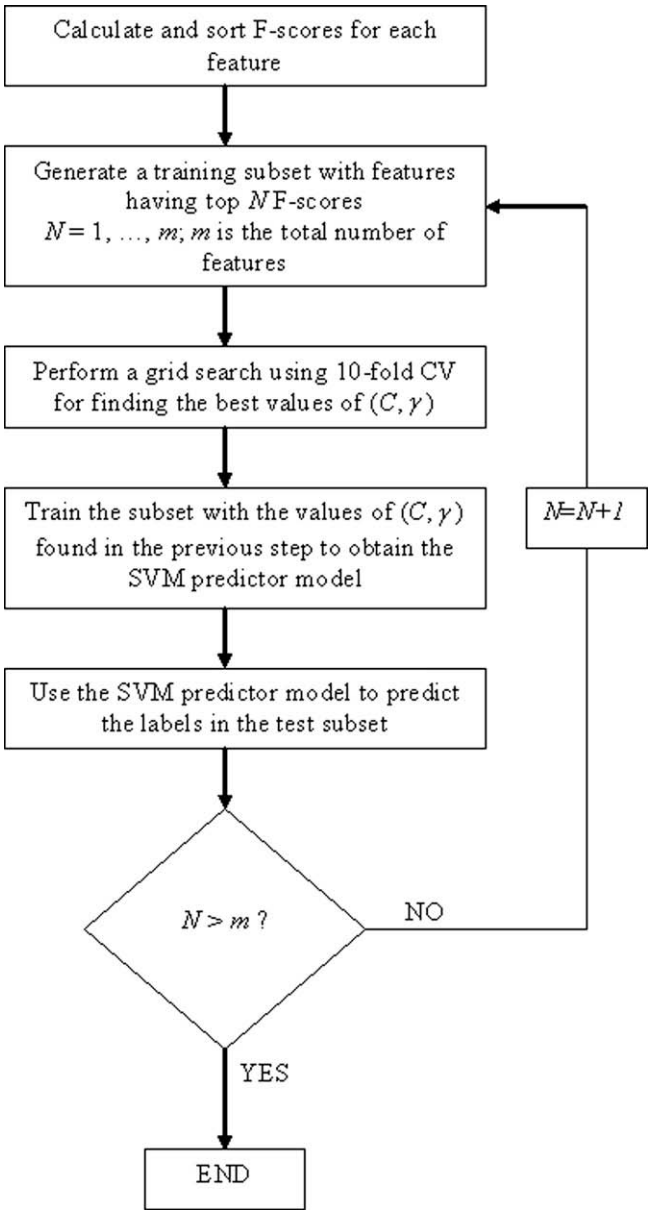

Fig. 1. SVM-based model using grid search to optimize model parameters and *F*-score calculation to select input features.

Table 1
Confusion matrix representation

Actual	Predicted	
	Positive	Negative
Positive	True positive (TP)	False negative (FN)
Negative	False positive (FP)	True negative (TN)

$$\text{Classification accuracy } (\%) = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}, \quad (20)$$

$$\text{Sensitivity } (\%) = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100, \quad (21)$$

$$\text{Specificity } (\%) = \frac{\text{TN}}{\text{FP} + \text{TN}} \times 100, \quad (22)$$

$$\text{Positive predictive value} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100, \quad (23)$$

$$\text{Negative predictive value} = \frac{\text{TN}}{\text{FN} + \text{TN}} \times 100. \quad (24)$$

ROC curves is a reliable technique based on the values of true positives and false positives, and therefore, provides a trade-off between sensitivity and specificity.

5. Results and discussion

To evaluate the effectiveness our method, we conducted experiments on the WBCD. The importance of each feature is measured by *F*-score, and the SVM parameters are optimized by grid search. Table 2 shows the relative importance with *F*-score for each feature on different training sets. The degree of breast tumor associated with features, from high to low, are F_6, F_3, F_2, F_1, F_7, F_8, F_5, F_4, and F_9. Therefore, we construct nine models with a different number of features to obtain the SVM classification models. Table 3 shows the nine models with different feature subsets based on *F*-score.

Table 4 shows the classification accuracies on the testing data for the nine models. Among the nine models, model #5 achieved the highest classification accuracy; 98.53% for the 50–50% training-test partition, 99.02% for the 70–30% training-test partition, and 99.51% for the 80–20% training-test partition. For comparison purposes, Table 5 gives the classification accuracies of our method and previous methods. As we can see from the results, our method using *F*-score and SVM obtains the highest classification accuracy so far.

We present values of sensitivity, specificity, positive predictive value and negative predictive value for model #5 in Table 6. The ROC curves for model #5 are also presented (Figs. 2–4). The areas under the ROC curves is computed,

Table 2
The relative feature importance with *F*-score

Feature no.	50–50% training-test partition	70–30% training-test partition	80–20% training-test partition
1	1.001444	1.133288	1.167029
2	1.328275	1.607558	1.786705
3	1.836976	1.826187	1.995388
4	1.798894	0.814200	0.866999
5	0.861221	0.834406	0.885195
6	1.990531	2.094498	2.192387
7	0.916463	1.066493	1.102302
8	0.897005	1.009641	1.054659
9	0.259275	0.277578	0.252970

Table 3
The nine feature subsets based on *F*-score

Model	No. of selected features	Features
#1	1	F_6
#2	2	F_6, F_3
#3	3	F_6, F_3, F_2
#4	4	F_6, F_3, F_2, F_1
#5	5	F_6, F_3, F_2, F_1, F_7
#6	6	F_6, F_3, F_2, F_1, F_7, F_8
#7	7	F_6, F_3, F_2, F_1, F_7, F_8, F_5
#8	8	F_6, F_3, F_2, F_1, F_7, F_8, F_5, F_4
#9	9	F_6, F_3, F_2, F_1, F_7, F_8, F_5, F_4, F_9

Table 4
Classification accuracies for each model and different test subsets

Model	Classification accuracy (%)		
	50–50% training-test partition	70–30% training-test partition	80–20% training-test partition
#1	92.10	91.21	91.21
#2	97.36	96.09	97.56
#3	97.95	98.04	97.56
#4	98.24	98.04	99.51
#5	98.53	99.02	99.51
#6	98.24	98.53	99.02
#7	98.24	98.53	98.53
#8	97.95	98.53	98.53
#9	98.24	98.53	99.02

Table 5
Classification accuracies obtained with our method and other classifiers from literature

Author (year)	Method	Classification accuracy (%)
Quinlan (1996)	C4.5	94.74
Hamiton et al. (1996)	RIAC	95.00
Ster and Dobnikar (1996)	LDA	96.80
Nauck and Kruse (1999)	NEFCLASS	95.06
Pena-Reyes and Sipper (1999)	Fuzzy-GA1	97.36
Setiono (2000)	Neuro-rule 2a	98.10
Goodman et al. (2002)	Optimized-LVQ	96.70
Goodman et al. (2002)	Big LVQ	96.80
Goodman et al. (2002)	AIRS	97.20
Albrecht et al. (2002)	LSA with perceptron algorithm	98.80
Abonyi and Szeifert (2003)	Supervised fuzzy clustering	95.57
Polat and Gunes (2007)	LS-SVM	98.53
This study (2007)	*F*-score + SVM	99.51

Table 6
Sensitivity, specificity, positive predictive value and negative predictive value for model #5

Measures	50–50% training-test partition	70–30% training-test partition	80–20% training-test partition
Sensitivity (%)	99.55	99.24	100
Specificity (%)	96.64	98.61	97.91
Positive predictive value (%)	98.22	99.24	98.88
Negative predictive value (%)	99.14	98.61	100

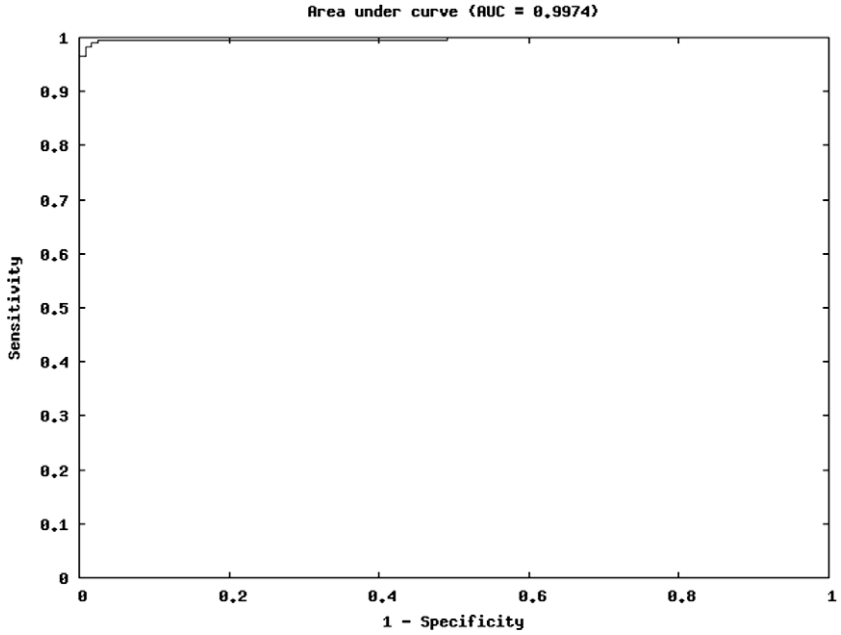

Fig. 2. ROC curve for 50–50% training-test partition.

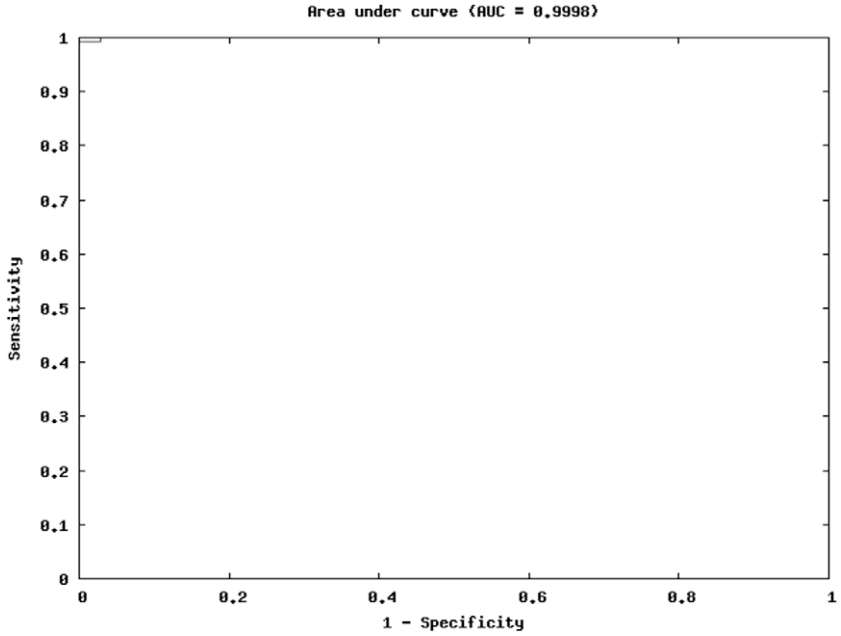

Fig. 3. ROC curve for 70–30% training-test partition.

and these values can be used for evaluating the classifier performance for different training/test partitions. The bigger area means better classifier performance.

In this study, there were two classes as benign and malignant. Classification results are displayed using a confusion matrix in Table 7. As we can see from Table 7, number of false positives and false negatives decrease with the increase of the training set size.

From the results above, we conclude that SVM with feature selection using *F*-score obtains promising results in classifying the potential breast cancer patients.

6. Conclusions

A medical decision making system based on SVM combined with feature selection has been applied on the task of diagnosing breast cancer. Experiments have been conducted on different portions of the WBCD, which is commonly used among researchers who use machine learning methods to diagnose breast cancer. It is observed that the proposed method yields the highest classification accuracies (98.53%, 99.02%, and 99.51% for 50–50% of training-test partition, 70–30% of training-test partition,

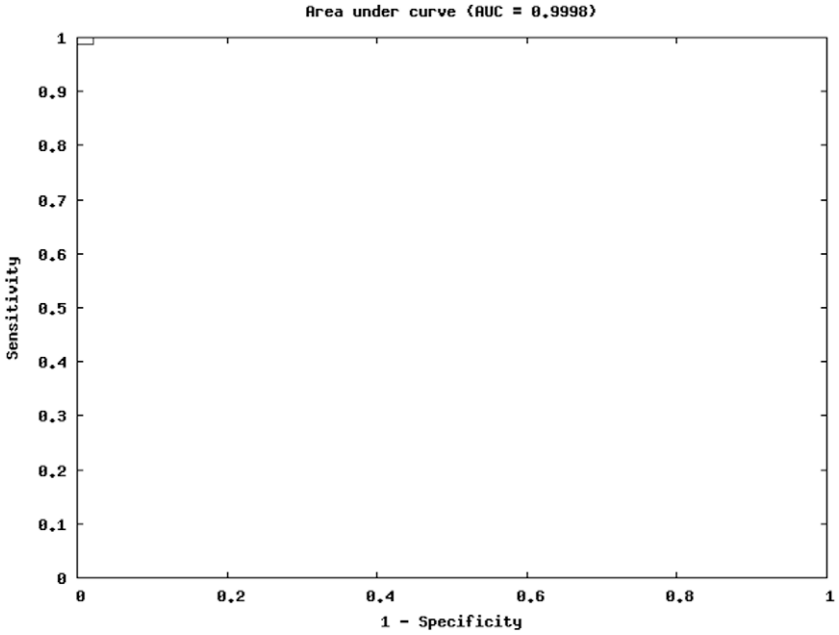

Fig. 4. ROC curve for 80–20% training-test partition.

Table 7
Confusion matrixes for model #5

Actual	Predicted		Partitions
	Benign	Malignant	
Benign	221	1	50–50% training-test partition
Malignant	4	115	
Benign	132	1	70–30% training-test partition
Malignant	1	71	
Benign	89	0	80–20% training-test partition
Malignant	1	47	

and 80–20% of training-test partition, respectively) for a subset that contained five features (model #5). Additional performance measures such as sensitivity, specificity, positive predictive value, negative predictive value, ROC curves and confusion matrices are also presented for model #5. Considering the results, the SVM-based model we have developed gives very promising results in classifying the breast cancer. We believe that the proposed system can be very helpful to the physicians for their final decisions on their patients. By using such a tool, they can make very accurate decisions.

Further exploration of the data can yield more interesting results. This will be the focus of our future work.

References

Abonyi, J., & Szeifert, F. (2003). Supervised fuzzy clustering for the identification of fuzzy classifiers. *Pattern Recognition Letters, 14*(24), 2195–2207.

Albrecht, A. A., Lappas, G., Vinterbo, S. A., Wong, C. K., & Ohno-Machado, L. (2002). Two applications of the LSA machine. In *Proceedings of the 9th international conference on neural information processing* (pp. 184–189).

Bertsekas, D. P. (1995). *Nonlinear programming*. Belmont: Athena Scientific.

Chen, Y. W., & Lin, C. J. (2005). Combining SVMs with various feature selection strategies. Available from http://www.csie.ntu.edu.tw/~cjlin/papers/features.pdf.

Courant, R., & Hilbert, D. (1953). *Methods of mathematical physics*. New York: Wiley Interscience.

Frohlich, H., & Chapelle, O. (2003). Feature selection for support vector machines by means of genetic algorithms. In *Proceedings of the 15th IEEE international conference on tools with artificial intelligence, Sacramento, CA* (pp. 142–148).

Goodman, D. E., Boggess, L., & Watkins, A. (2002). Artificial immune system classification of multiple-class problems. In *Proceedings of the artificial neural networks in engineering* (pp. 179–183).

Hamiton, H. J., Shan, N., & Cercone, N. (1996). RIAC: A rule induction algorithm based on approximate classification. Technical Report CS 96-06, University of Regina.

Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification. Available from http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *Proceedings of international conference machine learning*. Slovenia.

Kohavi, R., & Provost, F. (1998). Glossary of terms. *Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process, 30*(2–3).

Nauck, D., & Kruse, R. (1999). Obtaining interpretable fuzzy classification rules from medical data. *Artificial Intelligence in Medicine*(16), 149–169.

Osuna, E., Freund, R., & Girosi, F. (1997). Training support vector machines: Application to face detection. In *Proceedings of computer vision and pattern recognition, Puerto Rico* (pp. 130–136).

Pena-Reyes, C. A., & Sipper, M. (1999). A fuzzy-genetic approach to breast cancer diagnosis. *Artificial Intelligence in Medicine*(17), 131–155.

Polat, K., & Gunes, S. (2007). Breast cancer diagnosis using least square support vector machine. *Digital Signal Processing, 17*(4), 694–701.

Pontil, M., & Verri, A. (1998). Support vector machines for 3-D object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*(20), 637–646.

Quinlan, J. R. (1996). Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*(4), 77–90.

Scholkopf, B., Kah-Kay, S., Burges, C. J., Girosi, F., Niyogi, P., Poggio, T., et al. (1997). Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Transactions on Signal Processing, 45*(11), 2758–2765.

Setiono, R. (2000). Generating concise and accurate classification rules for breast cancer diagnosis. *Artificial Intelligence in Medicine, 18*(3), 205–217.

Ster, B., & Dobnikar, A. (1996). Neural networks in medical diagnosis: Comparison with other methods. In *Proceedings of the international conference on engineering applications of neural networks* (pp. 427–430).

Vapnik, V. (1989). *Statistical Learning Theory*. New York: Wiley.

Wan, V. & Campbell, W.M. (2000). Support vector machines for speaker verification and identification. Proceedings of IEEE Workshop Neural Networks for Signal Processing (pp. 775–784). Sydney.

West, D., Mangiameli, P., Rampal, R., & West, V. (2005). Ensemble strategies for a medical diagnosis decision support system: A breast cancer diagnosis application. *European Journal of Operational Research*(162), 532–551.

Zhang, G. P. (2000). Neural networks for classification: A survey. *IEEE Transactions on Systems, Man, and Cybernetics – Part C. Applications and Reviews, 30*(4), 451–462.