BREAST CANCER DIAGNOSIS AND PROGNOSIS USING DIFFERENT KERNEL-BASED CLASSIFIERS

Tingting Mu and Asoke K. Nandi
Department of Electrical Engineering and Electronics, The University of Liverpool, Brownlow Hill, Liverpool, UK, L69 3GJ

Keywords: Breast cancer, diagnosis, prognosis, pattern classification, kernel method.

Abstract: The medical applications of several advanced, kernel-based classifiers to breast cancer diagnosis and prognosis are studied and compared in this paper, including kernel Fisher's discriminative analysis, support vector machines (SVMs), multisurface proximal SVMs, as well as the pairwise Rayleigh quotient classifier and the strict 2-surface proximal classifier that we recently proposed. The radial basis function kernel is employed to incorporate nonlinearity. Studies are conducted with the Wisconsin diagnosis and prognosis breast cancer datasets generated from fine-needle-aspiration samples by image processing. Comparative analysis is provided in terms of classification accuracy, computing time, and sensitivity to the regularization parameters for the above classifiers.

1 INTRODUCTION

Despite the increasing public awareness and scientific research, breast cancer continues to be the most common form of cancer and the second most common cause of cancer deaths in females; the disease affects approximately 10% of all women at some stage of their life in the western world (Marshall, 1993). The long-term survival of a patient with breast cancer are improved by the early detection of the disease, which is enhanced by an accurate diagnosis. The choice of appropriate treatments following surgery is influenced by the expected long-term behavior of the disease, so-called prognosis.

Definitive diagnosis of a breast mass can only be established through fine-needle aspiration (FNA) biopsy, core needle biopsy, or excisional biopsy. Among these methods, FNA is the easiest and fastest method of obtaining a breast biopsy, and is effective for women who have fluid-filled cysts. Research works on the Wisconsin Diagnosis Breast Cancer (WDBC) data grew out of the desire of Dr. Wolberg to diagnose breast masses accurately based solely on FNA (Wolberg et al., 1993; Street et al., 1993). Later, a number of research projects have been developed with the WDBC dataset, focusing on computer-aided diagnosis (CAD) using machine learning techniques (Wolberg et al., 1994; Wolberg et al., 1995; Mangasarian et al., 1995; Guo and Nandi, 2006; Mu and Nandi, 2007). Breast cancer prognosis is a more difficult problem, that is, the long-term outlook for the disease for patients whose cancer has been surgically removed. Till now, few works have been developed on predicting the time to recur (TTR) for a patient for whom cancer has not recurred and may never recur (Wolberg et al., 1995; Mangasarian et al., 1995; Street et al., 1995). The detection of malignant breast tumors from a set of benign and malignant samples for diagnosis, and the simple prediction of patients as 'recurred' or 'not recurred' without predicting the TTR for prognosis, both belong to the pattern classification problems.

The idea of using kernel functions as inner product in a feature space was introduced into machine learning in 1964 by the work of Aizerman, Braverman and Rozonoer (Aizerman et al., 1964). Kernel methods to pattern analysis embeds the data in a suitable feature space, and then uses algorithms based on linear algebra, geometry, and statistics to discover patterns in the embedded data. Different kernel-based classifiers have been proposed. Boser, Guyon, and Vapnik (Boser et al., 1992) first combined the kernel function with the large margin hyperplanes, leading to support vector machines (SVMs) that are highly successful in solving various nonlinear and non-separable problems in machine learning. In addition to the original C-SVM learning method (Cortes and Vapnik, 1995), the ν-SVM learning method was proposed by Schölkopf et al. (Schölkopf et al., 2000), which is closely related to the C-SVM but with a different optimization risk. The famous Fisher's linear discriminant analysis (FLDA), dating back to 1936

Mu T. and K. Nandi A. (2008).
BREAST CANCER DIAGNOSIS AND PROGNOSIS USING DIFFERENT KERNEL-BASED CLASSIFIERS.
In *Proceedings of the First International Conference on Bio-inspired Systems and Signal Processing*, pages 342-348
DOI: 10.5220/0001059303420348

(Fisher, 1936), seeks separating hyperplanes which best separate two or more classes of samples based on the Fisher criterion with the between- and within-class scatters built on individual samples. Mika et al. (Mika et al., 1999) combined kernels functions with FLDA leading to kernel Fisher's discriminant analysis (KFDA). Mu et al. (Mu et al., 2007a) proposed to seek the optimal separating hyperplane based on the pairwise Rayleigh quotient (PRQ) criterion with the between- and within- class scatters built on the pairwise information; they also proposed to combine kernels functions with the linear PRQ classifier leading to the nonlinear PRQ classifier. Multiplane learning is a comparatively new machine learning method developed in recent years. Mangasarian and Wild (Mangasarian and Wild, 2006) proposed the kernel-based multisurface proximal SVM (MPSVM) that seeks two cross proximal planes by optimizing a regularized optimization objective with Tikhonov regularization term employed. More recently, Mu et al. (Mu et al., 2007b) proposed the strict 2-surface proximal (S2SP) classifier that seeks two cross proximal planes by employing a "square of sum" optimization factor without any regularization term, which is mathematically stricter than the optimization objective of MPSVM; and kernel functions were employed to incorporate nonlinearity.

In this paper, studies are conduced on the WDBC and WPBC datasets to investigate the benefits of applying different kernel-based classifiers to breast cancer diagnosis and prognosis, including SVM, KFDA, PRQ classifier, MPSVM, regularized δ-MPSVM (Mangasarian and Wild, 2006), and S2SP classifier. The detecting accuracies, computing times, and sensitivities to regularization parameters are compared for the above kernel-based classifiers.

2 CLASSIFICATION METHODS

Given a set of l labeled training samples $z = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{l} \in (R^n \times Y)$, where R^n is the n-dimensional real feature space with a binary label space $Y = \{1, -1\}$, and $y_i \in Y$ is the label assigned to the sample $\boldsymbol{x}_i \in R^n$, the purpose of classification is to seek the best prediction of the label for an input sample $\boldsymbol{x}$. All the kernel-based classifiers are developed in the kernel-transformed feature space κ, with a nonlinear mapping $\phi : R^n \rightarrow \kappa$.

2.1 Discriminant Classification

The basic idea of the discriminant classification is to seek one optimal hyperplane that best separates the two classes of samples in a corresponding feature space. In the kernel-transformed feature space κ, by expanding the direction vector of the hyperplane into a linear summation of all training samples, the separating hyperplane can be given as

$$f(\boldsymbol{x}) = \sum_{i=1}^{l} \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b, \tag{1}$$

where $\{\alpha_i\}_{i=1}^{l}$ denote the summating weights, b denotes the bias of the separating hyperplane, and $K(\cdot,\cdot)$ is a kernel function used to compute the inner product matrix, the so-called kernel matrix, on pairs of samples in the kernel-transformed feature space κ. Different classification methods lead to different ways to determine the optimal separating hyperplane $f^*(\boldsymbol{x})$. The label of a given test sample $\boldsymbol{x}$ can be predicted by

$$p(\boldsymbol{x}) = \mathrm{sgn}(f^*(\boldsymbol{x})), \tag{2}$$

where $\mathrm{sgn}(x)$ is equal to 1 when $x \geq 0$, and -1 otherwise.

2.1.1 Support Vector Machines

The basic idea of SVMs is to construct a separating hyperplane as the decision surface in such a way that the margin of separation between the positive and negative samples is maximized in an appropriate feature space. To determine $f^*(\boldsymbol{x})$ based on the maximal margin rule, the following constrained quadratic programming problem is solved (Cortes and Vapnik, 1995), as

$$O(\boldsymbol{\beta}) = \sum_{i=1}^{l} \beta_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y_i y_j \beta i \beta_j K(\boldsymbol{x}_i, \boldsymbol{x}_j), \tag{3}$$

subject to

$$\sum_{i=1}^{l} y_i \beta_i = 0,$$

$$0 \leq \beta_i \leq C, i = 1, 2, \ldots, l,$$

where $\{\beta_i\}_{i=1}^{l}$ are Lagrange multipliers, and C is the regularization parameter set by the user. Letting $\{\beta_i^*\}_{i=1}^{l}$ denote the optimal solution of $O(\boldsymbol{\beta})$, the optimal value of the summating weights $\{\alpha_i\}_{i=1}^{l}$ and the bias b are obtained by

$$\alpha_i^* = y_i \beta_i^*, i = 1, 2, \ldots, l, \tag{4}$$

$$b^* = -\frac{1}{2S} \sum_{\boldsymbol{x} \in S_+ \bigcup S_-} \sum_{i=1}^{l} y_i \beta_i^* K(\boldsymbol{x}, \boldsymbol{x}_i), \tag{5}$$

where S_+ and S_- are two sets of support vectors with the same size of S but different labels of $+1$ and -1.

2.1.2 Kernel Fisher's Discriminant Analysis

KFDA determines $f^*(\boldsymbol{x})$ by maximizing the following Fisher criterion (Shawe-Taylor and Cristianini, 2004), as

$$O(f) = \frac{(\mu^+ - \mu^-)^2}{(\sigma^+)^2 + (\sigma^-)^2}, \tag{6}$$

where

$$\mu^+ = \frac{1}{l^+}\left(\sum_{i=1}^{l^+} f(\boldsymbol{x}_i)\right)^2,$$

$$\mu^- = \frac{1}{l^-}\left(\sum_{i=1}^{l^-} f(\boldsymbol{x}_i)\right)^2,$$

$$(\sigma^+)^2 = \frac{1}{l^+}\sum_{i=1}^{l^+}(f(\boldsymbol{x}_i) - \mu^+)^2,$$

$$(\sigma^-)^2 = \frac{1}{l^-}\sum_{i=1}^{l^-}(f(\boldsymbol{x}_i) - \mu^-)^2,$$

where μ^+ and μ^- denote the mean projections of the positive and negative samples, respectively; σ^+ and σ^- are the corresponding standard deviations; and l^+ and l^- denote the number of samples from the positive and negative classes, respectively. By incorporating Eq. (1) into Eq. (6), the optimal values of $\{\alpha_i\}_{i=1}^{l}$ and b can be calculated by solving a generalized eigenvalue problem (Shawe-Taylor and Cristianini, 2004).

2.1.3 Pairwise Rayleigh Quotient Classifier

The PRQ classifier helps in classification with insufficient training samples by employing pairwise constraints instead of individual samples. To determine the optimal separating hyperplane $f^*(\boldsymbol{x})$, the following PRQ criterion is maximized (Mu et al., 2007a), as

$$O(f) = \frac{\tilde{d}}{\tilde{d}^+ + \tilde{d}^-}, \tag{7}$$

where

$$\tilde{d} = \left[\sum_{i=1}^{m}\frac{1}{2}(1 - z_i)(f_{i1} - f_{i2})\right]^2,$$

$$\tilde{d}^+ = \frac{1}{l^+(l^+ - 1)}\sum_{i=1}^{m}\frac{1}{4}(1 + z_i)(1 + y_{i1})(f_{i1} - f_{i2})^2$$

$$\tilde{d}^- = \frac{1}{l^-(l^- - 1)}\sum_{i=1}^{m}\frac{1}{4}(1 + z_i)(1 - y_{i1})(f_{i1} - f_{i2})^2,$$

where $\tilde{d}$ denotes the differences of projections between samples from different classes; $\tilde{d}^+$ denotes the differences of projections between samples from the positive class; $\tilde{d}^-$ denotes the differences of projections between samples from the negative class; y_{i1} denotes the label of the sample $\boldsymbol{x}_{i1}$; $z_i \in \{1, -1\}$ is the pairwise constraint assigned to the two samples in the pair $(\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2})$, and $z_i = 1$ if the two samples $(\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2})$ belong to the same class, whereas $z_i = -1$ if the two samples $(\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2})$ belong to different classes; f_{i1} and f_{i2} are used to denote $f(\boldsymbol{x}_{i1})$ and $f(\boldsymbol{x}_{i2})$; and m is the total number of available pairwise constraints. By incorporating Eq. (1) into Eq. (7), the optimal values of $\{\alpha_i\}_{i=1}^{l}$ and b can be simply calculated by matrix computation (Mu et al., 2007a). Compared with the Fisher criterion built on individual samples from a total number of l available samples, the PRQ criterion offers more possibilities by employing pairwise constraints from a total number of $l * (l - 1)$ available constraints.

2.2 Proximal Classification

The basic idea of proximal classification is to seek two proximal planes in a corresponding feature space, so that the first plane is as close to the points of the positive class while being as far as possible from the points of the negative class, whereas the second plane is as close to the points of the negative class while being as far as possible from the points of the positive class. In the kernel-transformed feature space κ, by expanding the direction vector of the hyperplane into a linear summation of all training samples, the two proximal hyperplanes are given as

$$f_1(\boldsymbol{x}) = \sum_{i=1}^{l}\alpha_{i1}K(\boldsymbol{x}_i, \boldsymbol{x}) + b_1, \tag{8}$$

$$f_2(\boldsymbol{x}) = \sum_{i=1}^{l}\alpha_{i2}K(\boldsymbol{x}_i, \boldsymbol{x}) + b_2, \tag{9}$$

where the subscripts 1 and 2 denote the first and second proximal plane, respectively. Let d_1 and d_2 denote the Euclidean distance between the sample and the two proximal planes, respectively, in the feature space κ. The label of a given test sample $\boldsymbol{x}$ can be predicted by considering values of d_1, d_2, and $\frac{d_1}{d_2}$ together using linear discriminant analysis.

2.2.1 Multisurface Proximal SVMs

MPSVMs obtain the first proximal hyperplane by maximizing the following objective function, as (Mangasarian and Wild, 2006)

$$O_1(\boldsymbol{\alpha}_1, b_1) = \frac{\|\mathbf{K}^-\boldsymbol{\alpha}_1 + \boldsymbol{e}b_1\|^2}{\|\mathbf{K}^+\boldsymbol{\alpha}_1 + \boldsymbol{e}b_1\|^2}; \tag{10}$$

and obtain the second proximal hyperplane by maximizing (Mangasarian and Wild, 2006)

$$O_2(\boldsymbol{\alpha}_2, b_2) = \frac{\|\mathbf{K}^+\boldsymbol{\alpha}_2 + \boldsymbol{e}b_2\|^2}{\|\mathbf{K}^-\boldsymbol{\alpha}_2 + \boldsymbol{e}b_2\|^2}, \tag{11}$$

where $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are two column vectors each with elements equal to $\{\alpha_{i1}\}_{i=1}^{l}$ and $\{\alpha_{i2}\}_{i=1}^{l}$, respectively; the $l^+ \times l$ matrix $\mathbf{K}^+$ represents the kernel matrix between the samples from the positive class and all the training samples; the $l^- \times l$ matrix $\mathbf{K}^-$ represents the kernel matrix between the samples from the negative class and all the training samples; and $\boldsymbol{e}$ is a column vector with all elements equal to one. The optimal values of $\boldsymbol{\alpha}_1$, b_1, $\boldsymbol{\alpha}_2$, and b_2 can be calculated by solving two generalized eigenvalue problems (Mangasarian and Wild, 2006), respectively.

Letting $\tilde{\boldsymbol{\alpha}_1}^T = \left[\boldsymbol{\alpha}_1^T, b_1\right]$, and $\tilde{\boldsymbol{\alpha}_2}^T = \left[\boldsymbol{\alpha}_2^T, b_2\right]$, to improve the classification performance of the MPSVMs, Mangasarian and Wild (Mangasarian and Wild, 2006) proposed to employ a Tikhonov regularization term, the two optimization objective shown in Eq. (10) and Eq. (11) become

$$O_1(\boldsymbol{\alpha}_1, b_1) = \frac{\|\mathbf{K}^-\boldsymbol{\alpha}_1 + \boldsymbol{e}b_1\|^2 + \delta\|\tilde{\boldsymbol{\alpha}_1}\|^2}{\|\mathbf{K}^+\boldsymbol{\alpha}_1 + \boldsymbol{e}b_1\|^2}, \tag{12}$$

$$O_2(\boldsymbol{\alpha}_2, b_2) = \frac{\|\mathbf{K}^-\boldsymbol{\alpha}_2 + \boldsymbol{e}b_2\|^2 + \delta\|\tilde{\boldsymbol{\alpha}_2}\|^2}{\|\mathbf{K}^+\boldsymbol{\alpha}_2 + \boldsymbol{e}b_2\|^2}, \tag{13}$$

where δ is a nonnegative regularization parameter set by the user. However, similar to the regularization parameter of the SVM, such as C for the C-SVM (Cortes and Vapnik, 1995) and ν for the ν-SVM (Schölkopf et al., 2000), performance of the above regularized δ-MPSVM is sensitive to the setting of the regularization parameter δ.

2.2.2 Strict 2-Surface Proximal Classifier

With consideration of the sign effect under the situation of misclassification with large projections onto the separating plane, the S2SP classifier eliminates the regularization term by employing the “square of sum” numerator. To obtain the first proximal hyperplane, the following objective function is to be maximized (Mu et al., 2007b), as

$$O_1(\boldsymbol{\alpha}_1, b_1) = \frac{\lceil \mathbf{K}^-\boldsymbol{\alpha}_1 + \boldsymbol{e}b_1 \rceil^2}{\|\mathbf{K}^+\boldsymbol{\alpha}_1 + \boldsymbol{e}b_1\|^2}, \tag{14}$$

and obtain the second proximal hyperplane by maximizing (Mu et al., 2007b)

$$O_2(\boldsymbol{\alpha}_2, b_2) = \frac{\lceil \mathbf{K}^+\boldsymbol{\alpha}_2 + \boldsymbol{e}b_2 \rceil^2}{\|\mathbf{K}^-\boldsymbol{\alpha}_2 + \boldsymbol{e}b_2\|^2}, \tag{15}$$

where $\lceil vector \rceil$ is used to denote the sum of the elements of the vector; and $\lceil matrix \rceil$ is used to denote a column vector with the sum of each row. The optimal values of $\boldsymbol{\alpha}_1$, b_1, $\boldsymbol{\alpha}_2$, and b_2 can be calculated by matrix computation (Mu et al., 2007b). There is no regularization parameter to be tuned for the S2SP classifier, which makes this method more convenient for the users, as compared with MPSVMs.

3 FEATURE PREPARATION

The WDBC and WPBC datasets were obtained from the University of Wisconsin Hospitals, Madison, of which the features were computed from digitized FNA samples. A portion of well-differentiated cells was scanned using a digital camera. The image analysis software system Xcyt was used to isolate individual nuclei (Wolberg et al., 1994; Wolberg et al., 1995; Mangasarian et al., 1995). In order to evaluate the size, shape, and texture of each cell nuclei, ten characteristics were derived and described as follows.

- **Radius** is computed by averaging the length of radial line segments from the center of mass of the boundary to each of the boundary points.
- **Perimeter** is measured as the sum of the distances between consecutive boundary points.
- **Area** is measured by counting the number of pixels on the interior of the boundary and adding one-half of the pixels on the perimeter, to correct for the error caused by digitization.
- **Compactness** combines the perimeter and area to give a measure of the compactness of the cell, calculated as $\frac{\text{perimeter}^2}{\text{area}}$.
- **Smoothness** is quantified by measuring the difference between the length of each radial line and the mean length of the two radial lines surrounding it, calculated by

  $$\frac{\sum_{\text{points}} |r_i - (r_i + r_{i+1})/2|}{\text{perimeter}},$$

  where r_i is the length of the line from the center of mass of the boundary to each boundary point.
- **Concavity** is captured by measuring the size of any indentations in the boundary of the cell nucleus.
- **Concave points** is similar to concavity, but counts only the number of boundary points lying on the concave regions of the boundary, rather than the magnitude of such concavities.
- **Symmetry** is 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, calculated by

$$\text{symmetry} = \frac{\sum_i |\text{left}_i - \text{right}_i|}{\sum_i (\text{left}_i + \text{right}_i)},$$

where left_i and right_i denote the lengths of perpendicular segments on the left and right of the major axis, respectively.

- **Fractal dimension** is approximated using the "coastline approximation" described by Mandelbrot (Mandelbrot, 1997). The perimeter of the nucleus is measured using increasingly larger "rulers". As the ruler size increases, the precision of the measurement decreases, and the observed perimeter decreases. Plotting these values on a log-log scale and measuring the downward slope gives the negative of an approximation to the fractal dimension.
- **Texture** is measured by finding the variance of the gray-scale intensities in the component pixels.

The mean value, standard error, and the extreme (largest or "worst") value of each characteristic were computed for each image, which resulted in 30 features of 569 images, yielding a database of 569×30 samples representing 357 benign and 212 malignant cases, for the WDBC dataset; and 30 features of 198 images, yielding a database of 198×30 samples representing 151 nonrecurring and 47 recurring cases, for the WPBC dataset.

4 EXPERIMENTS

Experiments and comparative analysis were conducted on the WDBC and WPBC datasets, using SVM, KFDA, PRQ classifier, MPSVM, regularized δ-MPSVM, and S2SP classifier. The features were normalized to have zero mean and unit variance before being used as the input of a classifier. Classification performance is shown in terms of classification accuracy in percentage. The radial basis function (RBF) kernel was employed to calculate the inner-product matrix between samples in the kernel-transformed feature space, given as

$$K(\boldsymbol{x}_a, \boldsymbol{x}_b) = \exp\left(-\frac{\|\boldsymbol{x}_a - \boldsymbol{x}_b\|^2}{2\sigma^2}\right),$$

where σ is the kernel width set by the user. The SVM was trained by using the "SVM and kernel methods MATLAB toolbox" (Canu et al., 2003).

The 10-fold-cross validation was used to evaluate the classifiers, which was executed by randomly dividing all the available samples into ten subsets each with nearly the same number of samples. The same ten sets of training-test trials were employed for every classification method, each with one subset for test and the remaining nine subsets for training. Parameters of each classifier were selected by using the 5-fold-cross validation within the training set of the first trial. The same five sets of training-test trials were conducted to select parameters for each classification method. Finally, the mean value of the ten test classification accuracies with the selected parameters was used to represent the generalized performance.

Table 1: Performance comparison in percentage accuracy and computing time for different kernel-based classifiers.

	WDBC		WPBC	
	Accu.	Time	Accu.	Time
Methods	(%)	(Sec.)	(%)	(Sec.)
SVM	98.8	0.09	76.3	0.08
KFDA	97.2	0.09	76.3	0.02
PRQ	97.7	8.02	76.3	1.07
MPSVM	85.3	0.90	75.3	0.21
δ-MPSVM	91.6	0.67	76.3	0.09
S2SP	99.2	0.10	77.3	0.02
Lam et al.	95.6	N/A	76.3	N/A

The classification performance and the corresponding computing time of each classifier are recorded in Table 1 using the WDBC and WPBC datasets; the results were also compared with the 10-fold-cross-validation performance obtained by the edited nearest-neighbor (ENN) with pure filtering (Lam et al., 2002) using the same datasets. The S2SP classifier provided the best classification accuracy of 99.2% as compared with the other five kernel-based classifiers. Nearly all of our obtained results (above 97%) were better than the published result of 95.6% (Lam et al., 2002) (see Table 1). For the more difficult WPBC dataset, the S2SP classifier provided the best classification accuracy of 77.3%. KFDA, SVM, δ-MPSVMs, and the PRQ classifier provided the same performance of 76.3% as that obtained by ENN (Lam et al., 2002) (see Table 1). KFDA, SVM, and the S2SP classifier possess faster training speed than MPSVMs, δ-MPSVMs, and the PRQ classifier, and performs better than MPSVMs and δ-MPSVMs. The classification performance of the PRQ classifier is comparable to those obtained by KFDA, SVM, and the S2SP classifier.

For a reasonable comparison of the classification capabilities, a score is calculated by averaging the classification performance over the two datasets and timing by 100 for each classifier, and recorded in Table 2. It can be seen from Table 2 that the S2SP classifier provides the highest score and requires

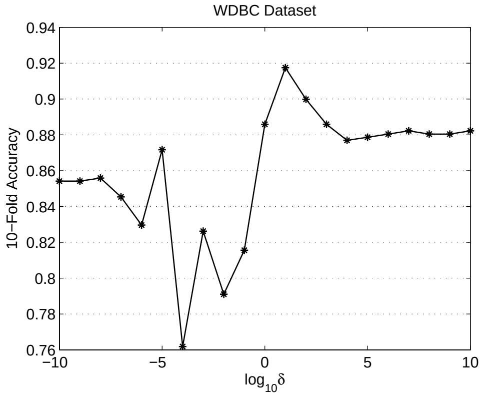

Figure 1: Performance variations of the δ-MPSVM classifier versus different values of $\log_{10..}$ δ, with the RBF kernel width σ fixed as the selected values, for the WDBC dataset.

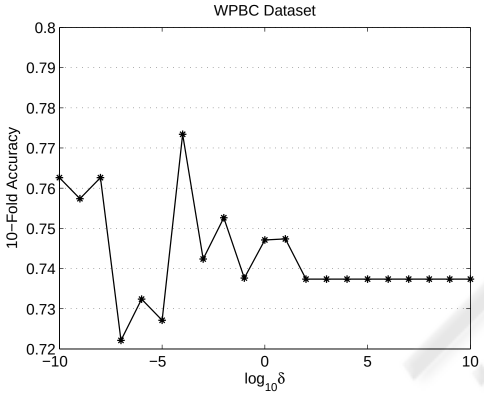

Figure 2: Performance variations of the δ-MPSVM classifier versus different values of $\log_{10..}$ δ, with the RBF kernel width σ fixed as the selected values, for the WPBC dataset.

the least parameters to be tuned. Both the SVM and δ-MPSVM classifiers require to determine one extra regularization parameter. For SVM, C controls the tradeoff between the complexity of a SVM and the number of non-separable points. The SVM performance is not very sensitive to the setting of C. The performance variations of the δ-MPSVM are provided in Fig. 1 and Fig. 2, by varying the value of $\log_{10}\delta$ from -10 to 10, for the WDBC and WPBC datasets, respectively. It can be seen from Fig. 1 and Fig. 2 that performance of the δ-MPSVM classifier is sensitive to the setting of δ. Without using the regularization term, the average score of the MPSVM classifier falls down from 84.0 to 80.3 (see Table 2). However, tuning of the values of the kernel parameters is unavoidable for all these kernel-based classifiers.

Table 2: Comparison of classification capability in average percentage accuracy for different classifiers.

Rank	Classifiers	Score	Parameters
1	S2SP	88.3	1 (σ)
2	SVM	87.6	2 (σ, C)
3	PRQ	87.0	1 (σ)
4	KFDA	86.8	1(σ)
5	δ-MPSVM	84.0	2 (σ, δ)
6	MPSVM	80.3	1 (σ)

5 CONCLUSIONS

Five recently developed, kernel-based, nonlinear classifiers, including SVM, KFDA, PRQ classifier, MPSVMs (unregularized MPSVM and regularized δ-MPSVM), and S2SP classifier, have been applied to breast cancer diagnosis and prognosis. We have studied and compared the benefits of the above classifiers in terms of classification accuracy, computing time, and sensitivity to the regularization parameter. Studies were conducted with the WDBC and WPBC datasets. Experimental results demonstrate that the classification accuracies of SVM, KFDA, S2SP, and PRQ classifiers are comparable. However, the PRQ classifier possesses the slowest computing speed, as the PRQ criterion built on pairwise constrains leads to an increase of the computing speed by l^2 as the size (l) of the training samples increases. The classification performance of MPSVM is unsatisfactory, and sensitive to the setting of the regularization parameter δ. From an overall consideration, the S2SP classifier is more favorable to users with not only higher classification accuracy but also faster computing speed; furthermore, there is no regularization parameter to be tuned for the S2SP classifier.

ACKNOWLEDGEMENTS

T. Mu would like to acknowledge financial support from the Overseas Research Students Awards Scheme (ORSAS), the Hsiang Su Coppin Memorial Scholarship Fund, and the University of Liverpool, UK. We thank the Medical Research Council (the Interdisciplinary Bridging Awards), UK, for financial support.

REFERENCES

Aizerman, M., Braverman, E., and Rozonoer, L. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proc. of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152.

Canu, S., Grandvalet, Y., and Rakotomam, A. (2003). *SVM and Kernel Methods Matlab Toolbox*. Perception Systems et Information, INSA de Rouen, Rouen, France.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.

Guo, H. and Nandi, A. K. (2006). Breast cancer diagnosis using genetic programming generated feature. *Pattern Recognition*, 39:980–987.

Lam, W., Keung, C., and Ling, C. X. (2002). Learning good prototypes for classification using filtering and abstraction of instances. *Pattern Recognition*, 35(7):1491–1506.

Mandelbrot, B. B. (1997). *The Fractal Geometry of Nature, Chapter 5*. W. H. Freeman and Company, New York.

Mangasarian, O. L., Street, W. N., and Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43(4):570–577.

Mangasarian, O. L. and Wild, E. W. (2006). Multisurface proximal support vector machine classification via generalized eigenvalues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:69–74.

Marshall, E. (1993). Search for a killer: Focus shifts from fat to hormones in special report on breast cancer. *Science*, 259:618–621.

Mika, S., Rätsch, G., Weston, J., Schölkopf, B., and Muller, K. (1999). Fisher discriminant analysis with kernels. In *Proc. of IEEE Neural Networks for Signal Processing Workshop*, pages 41–48.

Mu, T. and Nandi, A. K. (2007). Breast cancer detection from FNA using SVM with different parameter tuning systems and SOM–RBF classifier. *Journal of the Franklin Institute*, 344(3-4):285–311.

Mu, T., Nandi, A. K., and Rangayyan, R. M. (2007a). Pairwise Rayleigh quotient classifier with application to the analysis of breast tumors. In *Proc. of the 4th IASTED Int'l Conf. on Signal Processing, Pattern Recognition, and Applications, SPPRA*, pages 356–361, Innsbruck, Austria.

Mu, T., Nandi, A. K., and Rangayyan, R. M. (2007b). Strict 2-surface proximal classifier with application to breast cancer detection in mammograms. In *Proc. of the 32nd Int'l Conf. on Acoustics, Speech, and Signal Processing, ICASSP*, volume 2, pages 477–480, Honolulu, HI.

Schölkopf, B., Smola, A. J., Williamson, R., and Bartlett, P. (2000). New support vector algorithms. *Neural Computation*, 12:1207–1245.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK.

Street, W. N., Mangasarian, O. L., and Wolberg, W. H. (1995). An inductive learning approach to prognostic prediction. In *Proc. of the 12th Int'l Conf. on Machine Learning, ICML*, pages 522–530, Morgan Kaufmann.

Street, W. N., Wolberg, W. H., and Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. In *Proc. of IST/SPIE Symposium on Electronic Imaging: Science and Technology*, volume 1905, pages 861–870, San Jose, CA.

Wolberg, W. H., Street, W. N., and Mangasarian, O. L. (1993). Breast cytology diagnosis via digital image analysis. *Analytical and Quantitative Cytology and Histology*, 15(6):396–404.

Wolberg, W. H., Street, W. N., and Mangasarian, O. L. (1994). Machine learning techniques to diagnose breast cancer from fine-needle aspirates. *Cancer Letter*, 77:163–171.

Wolberg, W. H., Street, W. N., and Mangasarian, O. L. (1995). Image analysis and machine learning applied to breast cancer diagnosis and prognosis. *Analytical and Quantitative Cytology and Histology*, 17(2):77–87.