# STRICT 2-SURFACE PROXIMAL CLASSIFIER WITH APPLICATION TO BREAST CANCER DETECTION IN MAMMOGRAMS


*Tingting Mu, Asoke K. Nandi*

The University of Liverpool
Department of Electrical Engineering and Electronics
Brownlow Hill, L69 3GJ, Liverpool, UK

*Rangaraj M. Rangayyan*

University of Calgary
Department of Electrical and Computer Engineering
Schulich School of Engineering
Calgary, Alberta T2N 1N4, Canada



## ABSTRACT

We propose a 2-plane learning method for binary classification, named as the strict 2-surface proximal (S2SP) classifier, by seeking two cross proximal planes based on two strict optimization objectives with a "square of sum" optimization factor, of which the nonlinearity is achieved by employing kernel functions. We apply the S2SP classifier for both linear and nonlinear classification to recognize malignant tumors from a set of 57 regions in mammograms, of which 20 are related to malignant tumors and 37 to benign masses. Ten different feature combinations are studied. Experimental results demonstrate that the linear S2SP classifier provides results comparable to those obtained by Fisher linear discriminant analysis (FLDA). For one feature set ($FS_8$, see Table 2), the linear classification performance was significantly improved to $0.97$ by using the S2SP classifier, as compared to the FLDA performance of $0.82$, in terms of the area under the receiver operating characteristics (ROC) curve. In the case of nonlinear classification, the S2SP classifier with the triangle kernel provided a perfect performance of $1.0$ for all of the ten feature combinations, also evaluated in terms of the area under the ROC curve, but with good robustness limited to the setting of the kernel parameter in a certain range.



***Index Terms*—** Multiplane learning, proximal classification, square of sum, breast cancer, breast tumors


## 1. INTRODUCTION

Multiplane learning is a comparatively new machine learning method developed in recent years. Bradley and Mangasarian [1] addressed the topic of multiplane learning by proposing the unsupervised $k$-plane clustering method. Later, series of studies were conducted on multiplane learning for supervised pattern classification [2, 3, 4]. Proximal support vector machines (PSVMs) [2] generate two parallel planes such that each plane is closer to one of the two data sets to be classified while also being as far apart as possible. Pal et al. [3] proposed a fuzzy extension of PSVM via generalized eigenvalues. More recently, Mangasarian and Wild [4] dropped the parallel condition in the PSVM, which leads to the multisurface PSVM (MPSVM) via the solution of a generalized eigenvalue problem. In their study, a Tikhonov regularization term is employed to improve the performance of the MPSVM. However, MPSVM users face the problem of tuning the regularization parameter $\delta$ for the regularized $\delta$-MPSVM, of which the performance is sensitive to the setting of $\delta$.

In this work, we propose a strict 2-surface proximal (S2SP) learning method without any regularization term, which seeks two planes for binary classification by employing a "square of sum" optimization factor. Compared with the MPSVM classifier, the "square of sum" form leads to two stricter optimization objectives for the S2SP classifier. We apply the proposed method to recognize malignant tumors from a set of 57 regions in mammograms, of which 20 are related to malignant tumors and 37 to benign masses, and compare the results with those obtained by Fisher linear discriminant analysis (FLDA), as well as those obtained by André and Rangayyan [5] using artificial neural networks (ANNs), and Alto et al. [6] using linear discriminant analysis (LDA), to demonstrate the efficiency of the S2SP classifier.

## 2. THE S2SP CLASSIFIER

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 binary classification is to seek the best prediction of the label for an input sample $\boldsymbol{x}$. The basic idea of proximal classification [2, 4] is to seek two proximal planes in a corresponding feature space, i.e., $R^n$:

$$f_1(\boldsymbol{x}) = \boldsymbol{\omega}_1^T \boldsymbol{x} + b_1 = 0, \tag{1}$$

$$f_2(\boldsymbol{x}) = \boldsymbol{\omega}_2^T \boldsymbol{x} + b_2 = 0, \tag{2}$$

where $\boldsymbol{\omega}$ and $b$ are the weight vector (direction) and bias of the proximal planes, respectively; the subscripts 1 and 2 denote the first and second plane, respectively; the first plane is closest to the points of the positive class and farthest from the points of the negative class, whereas the second plane is closest to the points of the negative class and farthest from the points of the positive class.

For the S2SP classifier, we propose to seek the two proximal planes by maximizing two strict objectives with numerators in the "square of sum" form, given by

$$\frac{\left[\sum_{i=1}^{l}(1-y_i)f_1(\boldsymbol{x}_i)\right]^2}{\sum_{i=1}^{l}(1+y_i)f_1^2(\boldsymbol{x}_i)}, \tag{3}$$

$$\frac{\left[\sum_{i=1}^{l}(1+y_i)f_2(\boldsymbol{x}_i)\right]^2}{\sum_{i=1}^{l}(1-y_i)f_2^2(\boldsymbol{x}_i)}. \tag{4}$$

For linear classification, with $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$ given in Eq.(1) and Eq.(2), Eq.(3) leads to the following objective function to obtain the first proximal plane:

$$O_1'(\boldsymbol{\omega}, b) = \frac{\lceil X^-\boldsymbol{\omega} + \boldsymbol{e}b\rceil^2}{\|X^+\boldsymbol{\omega} + \boldsymbol{e}b\|^2}, \tag{5}$$

where $\boldsymbol{e}$ is a column vector with all elements equal to 1; $\lceil vector\rceil$ 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. To obtain the second proximal plane, Eq.(4) leads to the following objective function:

$$O_2'(\boldsymbol{\omega}, b) = \frac{\lceil X^+\boldsymbol{\omega} + \boldsymbol{e}b\rceil^2}{\|X^-\boldsymbol{\omega} + \boldsymbol{e}b\|^2}. \tag{6}$$

Letting $\boldsymbol{\varpi} = [\boldsymbol{\omega}, b]^T$ and $\tilde{X}^+ = \begin{bmatrix} X^+ & \boldsymbol{e} \end{bmatrix}$, $\tilde{X}^- = \begin{bmatrix} X^- & \boldsymbol{e} \end{bmatrix}$, the optimal solution of Eq.(5) and Eq.(6) can be calculated by solving two generalized eigenvalue problems based on two Rayleigh quotients derived from Eq.(5) and Eq.(6), respectively, given by

$$\boldsymbol{\varpi}_1^* = Q_1^{-1}\lceil\tilde{X}^-\rceil, \tag{7}$$

$$\boldsymbol{\varpi}_2^* = Q_2^{-1}\lceil\tilde{X}^+\rceil, \tag{8}$$

where

$$Q_1 = (\tilde{X}^+)^T\tilde{X}^+, \quad Q_2 = (\tilde{X}^-)^T\tilde{X}^-.$$

For nonlinear classification, in the transformed kernel feature space $\kappa$, by expanding the weight vectors of the planes into a linear summation of all training samples, the following two kernel-based proximal planes are employed:

$$f_1(\boldsymbol{x}) = \sum_{i=1}^{l}\alpha_{1,i}K(\boldsymbol{x}_i, \boldsymbol{x}) + b_1 = 0, \tag{9}$$

$$f_2(\boldsymbol{x}) = \sum_{i=1}^{l}\alpha_{2,i}K(\boldsymbol{x}_i, \boldsymbol{x}) + b_2 = 0, \tag{10}$$

where $K(\cdot,\cdot)$ is a kernel function used to compute the inner product matrix, the so-called kernel matrix, on pairs of samples in the transformed feature space $\kappa$. By incorporating Eq.(9) into Eq.(3), we get the following objective function to be maximized for the first plane:

$$O_{\kappa 1}'(\boldsymbol{\omega}, b) = \frac{\lceil K^-\boldsymbol{\alpha} + \boldsymbol{e}b\rceil^2}{\|K^+\boldsymbol{\alpha} + \boldsymbol{e}b\|^2}, \tag{11}$$

where the $l^+ \times l$ matrix $K^+$ represents the kernel matrix between the samples from the positive class and all the training samples, and the $l^- \times l$ matrix $K^-$ represents the kernel matrix between the samples from the negative class and all the training samples. For the second plane, by incorporating Eq.(10) into Eq.(4), the function to be maximized is

$$O_{\kappa 2}'(\boldsymbol{\omega}, b) = \frac{\lceil K^+\boldsymbol{\alpha} + \boldsymbol{e}b\rceil^2}{\|K^-\boldsymbol{\alpha} + \boldsymbol{e}b\|^2}. \tag{12}$$

Letting $\tilde{\boldsymbol{\alpha}} = [\boldsymbol{\alpha}, b]^T$ and adding a column with all elements equal to 1 to the kernel matrices of $K^+$ and $K^-$ as

$$\tilde{K}^+ = \begin{bmatrix} K^+ & \boldsymbol{e} \end{bmatrix}, \quad \tilde{K}^- = \begin{bmatrix} K^- & \boldsymbol{e} \end{bmatrix},$$

the optimal solution of Eq.(11) and Eq.(12) can be derived as

$$\tilde{\boldsymbol{\alpha}}_1^* = Q_{\kappa 1}^{-1}\lceil\tilde{K}^-\rceil, \tag{13}$$

$$\tilde{\boldsymbol{\alpha}}_2^* = Q_{\kappa 2}^{-1}\lceil\tilde{K}^+\rceil, \tag{14}$$

where

$$Q_{\kappa 1} = (\tilde{K}^+)^T\tilde{K}^+, \quad Q_{\kappa 2} = (\tilde{K}^-)^T\tilde{K}^-.$$

There are no extra parameters to be tuned, except for the kernel parameter for nonlinear classification, which is necessary for all kernel-based methods.

To seek the two proximal planes, the MPSVM classifier, proposed by Mangasarian and Wild [4], maximizes two objectives with the numerator parts given by $\sum_{i=1}^{l}(1-y_i)f^2(\boldsymbol{x}_i)$ that are in the "sum of squares" form, compared with the numerator parts in Eq.(3) and Eq.(4) that are in the "square of sum" form. Mangasarian and Wild [4] also introduced Tikhonov regularization terms into their optimization objectives to reduce the norm of $(\boldsymbol{\omega}, b)$, which efficiently improves the performance of the MPSVM classifier. For our proposed S2SP classifier, the optimal solution of the two proximal planes is provided by employing the "square of sum" numerator, with consideration of the sign effect under the situation of misclassification with large projections onto the separating plane. In the case of the MPSVM classifier, the performance is improved by employing the Tikhonov regularization term. However, the performance of the regularized $\delta$-MPSVM is sensitive to the setting of the regularization parameter $\delta$.

| Feature sets | Features used |
|---|---|
| $FS_1$ | $F_{cc}$, $A$ |
| $FS_2$ | $F_{cc}$ and $f_8$ |
| $FS_3$ | $F_{cc}$, $A$, and $f_8$ |
| $FS_4$ | $A$, $Co$, and $CV$ |
| $FS_5$ | $A$ and $f_8$ |
| $FS_6$ | 14 texture features |
| $FS_7$ | $F_{cc}$ and 14 texture features |
| $FS_8$ | $A$, $f_4$, $f_6$, $f_7$, $f_8$, $f_9$, and $f_{14}$ |
| $FS_9$ | $F_{cc}$, $f_3$, $f_4$, $f_5$, and $f_{12}$ |
| $FS_{10}$ | All 18 features |

**Table 1**. Different feature combinations used in the study.

## 3. EXPERIMENTS AND RESULTS

Fifty-seven regions in mammograms, of which 20 are related to malignant tumors and 37 to benign masses, are used in this study, obtained from "Screen Test: Alberta Program for the Early Detection of Breast Cancer" [6, 7]. Eighteen real-valued features are considered for each sample, including one shape feature known as fractional concavity ($F_{cc}$), three edge-sharpness features known as acutance ($A$), contrast ($Co$), and coefficient of variation ($CV$), as well as 14 texture features known as energy ($f_1$), contrast ($f_2$), correlation ($f_3$), sum of squares ($f_4$), inverse difference moment ($f_5$), sum average ($f_6$), sum variance ($f_7$), sum entropy ($f_8$), entropy ($f_9$), difference variance ($f_{10}$), difference entropy ($f_{11}$), information measure of correlation ($f_{12}$ and $f_{13}$), and maximal correlation coefficient ($f_{14}$) [6]. Ten different feature combinations, as listed in Table 1, were studied in this work, of which the feature combinations $FS_8$ and $FS_9$ were from the feature selection results using genetic algorithm. The nonlinearity of the S2SP classifier was achieved by employing the triangle kernel, given by

$$K(\boldsymbol{x}_a, \boldsymbol{x}_b) = \max\left(1 - \frac{\|\boldsymbol{x}_a - \boldsymbol{x}_b\|}{\sqrt{2}\sigma}, 0\right), \qquad (15)$$

where $\sigma$ is the kernel width set by the user. The leave-one-out (LOO) procedure was used to evaluate the classification performance for both linear and nonlinear classifiers, because of the small size of the dataset. All of the features were normalized before being classified. Classification performance is shown in terms of the area under the receiver operating characteristics (ROC) curve, named $A_z$, generated by applying a sliding threshold with the LS-SVMlab1.5 toolbox [8].

### 3.1. Linear Classification

In this experiment, the linear S2SP classifier without using any kernel function, as well as FLDA were applied on the ten different feature combinations as listed in Table 1, of which the corresponding performance in $A_z$ values is recorded in Table 2, and compared with those obtained by André and Rangayyan [5] using the single-layer perceptron, and by Alto et al. [6] using LDA. Proximal planes learned by the S2SP classifier as well as the decision boundary learned by FLDA with the feature set $FS_1$ are shown in the part (a) of Fig. 1. It can be seen from Table 2 that the linear classification performance obtained by the S2SP classifier is comparable to those obtained by FLDA, and better than those obtained by Alto et al. [6] and André and Rangayyan [5] for most feature combinations. It is worth mentioning that the linear classification performance of feature sets $FS_6$ and $FS_8$ in $A_z$ values has been significantly improved to $0.68$ and $0.97$ by using the S2SP classifier, respectively, which is much better than the FLDA performance of $0.63$ and $0.82$; for feature set $FS_9$, the perfect performance of $A_z = 1.0$ was reached. However, it should be remarked that the proximal classification method does not perform well for all cases, compared with the discriminative classification method. For feature set $FS_7$, the performance of the S2SP classifier is not satisfactory.

| | S2SP | | FLDA | | Others |
|---|---|---|---|---|---|
| Features | $A_z$ | SE | $A_{z1}$ | SE | $A_{z2}$ |
| $FS_1$ | 0.997 | 0.004 | 0.997 | 0.004 | 0.98 [6] |
| $FS_2$ | 0.997 | 0.004 | 0.999 | 0.002 | 0.99 [6] |
| $FS_3$ | 0.997 | 0.004 | 0.997 | 0.004 | 0.99 [6] |
| $FS_4$ | 0.71 | 0.08 | 0.72 | 0.09 | 0.62 [5] |
| $FS_5$ | 0.76 | 0.06 | 0.79 | 0.07 | 0.76 [6] |
| $FS_6$ | ***0.68*** | 0.07 | 0.63 | 0.08 | 0.65 [5] |
| $FS_7$ | 0.79 | 0.05 | 0.99 | 0.008 | N/A |
| $FS_8$ | ***0.97*** | 0.02 | 0.82 | 0.07 | N/A |
| $FS_9$ | 1.0 | 0.0 | 0.99 | 0.006 | N/A |
| $FS_{10}$ | 0.94 | 0.03 | 0.95 | 0.03 | N/A |

**Table 2**. Linear classification performance in $A_z$ values for different feature combinations using the S2SP classifier in the original feature space ($A_z$), and the results of FLDA ($A_{z1}$) as well as the results obtained by André and Rangayyan [5] and Alto et al. [6] ($A_{z2}$) for comparison. SE denotes the standard error. Significantly improved $A_z$ values are shown in bold.

### 3.2. Nonlinear Classification

In this experiment, the S2SP classifier with the triangle kernel given in Eq.(15) was applied to the same ten feature combinations as listed in Table 1. It was found that all the ten feature combinations could reach the perfect performance of $A_z = 1.0$ by employing the triangle kernel with an appropriate kernel width $\sigma$ for the S2SP classifier. The selected values of the kernel parameter $\sigma$ to reach the perfect performance of $A_z = 1.0$, for the ten feature combinations, are recorded in Table 3. However, it was observed that most of the feature combinations are robust to variation of $\sigma$ in certain limited intervals. We show a plot of the variation of the nonlinear classification performance in $A_z$ values versus dif-

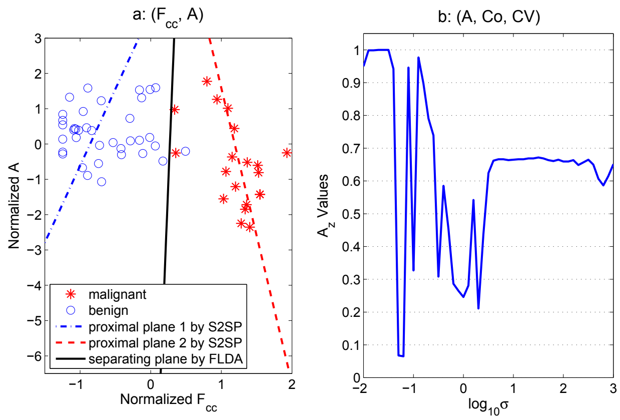


**Fig. 1**. (a): Proximal planes of the S2SP classifier and the decision boundary of FLDA. (b): Variation of the nonlinear classification performance.

| | $FS_1$ | $FS_2$ | $FS_3$ | $FS_4$ | $FS_5$ |
|---|---|---|---|---|---|
| $\log_{10}\sigma$ | -1.5 | -2.3 | -1.4 | -1.6 | -1.7 |
| | $FS_6$ | $FS_7$ | $FS_8$ | $FS_9$ | $FS_{10}$ |
| $\log_{10}\sigma$ | -2.8 | -1.7 | -2.1 | -1.9 | -2.3 |

**Table 3**. Parameter settings of $\sigma$ for different feature sets, using the S2SP classifier with the triangle kernel, in order to obtain $A_z = 1.0$.

ferent values of $\log_{10}\sigma$ for the feature set $FS_4$ by using the S2SP classifier with the triangle kernel in part (b) of Fig. 1. It can be seen that the $A_z$ value varies significantly for different values of $\log_{10}\sigma$, especially when $\log_{10}\sigma$ is less than 0. Similar results were observed for the other feature combinations. Further study is in progress on the evaluation of the robustness around the selected kernel parameter values, and on optimization of the kernel parameter itself.

## 4. CONCLUSION

We have proposed the S2SP classifier for both linear and nonlinear pattern classification. The classifier seeks two proximal planes in a corresponding feature space by maximizing two strict optimization objectives with the "square of sum" optimization term. The S2SP classifier was applied to identify malignant tumors from a set of 57 regions in mammograms, of which 20 are related to malignant tumors and 37 to benign masses. Experimental results demonstrate the effectiveness of the linear S2SP classifier in terms of improved performance as compared to that obtained by FLDA. More importantly, with the nonlinear S2SP classifier, by employing the triangle kernel with an appropriate kernel width $\sigma$, all of the feature combinations tested could achieve the perfect performance of $A_z = 1.0$, but with instability to variation of $\sigma$. The S2SP classifier has shown promise in improving the accuracy of discriminating between benign breast masses and malignant tumors based on features that provided weak performance using classical pattern recognition methods. The proposed methods should find application in computer-aided detection and diagnosis of breast cancer.

## 5. ACKNOWLEDGMENT

T. Mu would like to acknowledge financial support from the ORSAS, UK; the Hsiang Su Coppin Memorial Scholarship Fund and the University of Liverpool, UK. We also thank the University Research Grants Committee of the University of Calgary, Canada, as well as the Medical Research Council (the Interdisciplinary Bridging Awards), UK, for financial support.

## 6. REFERENCES

[1] P. S. Bradley and O. L. Mangasarian, "k-plane clustering," *Journal of Global Optimization*, vol. 16, pp. 23–32, Jan. 2000.

[2] G. M. Fung and O. L. Mangasarian, "Multicategory proximal support vector machine classifiers," *Machine Learning*, vol. 59, pp. 77–97, May 2005.

[3] S. K. Pal, S. Bandyopadhyay, and S. Biswas, "Fuzzy proximal support vector classification via generalized eigenvalues," in *Proc. of 1st Int'l Conf. on Pattern Recognition and Machine Intelligence*, Kolkata, India, 2005, pp. 360–363.

[4] O. L. Mangasarian and E. W. Wild, "Multisurface proximal support vector machine classification via generalized eigenvalues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 69–74, Jan. 2006.

[5] T. C. S. S. André and R. M. Rangayyan, "Classification of breast masses in mammograms using neural networks with shape, edge sharpness, and texture features," *Journal of Electronic Imaging*, vol. 15, 2005, article no. 013019.

[6] H. Alto, R. M. Rangayyan, and J. E. L. Desautels, "Content-based retrieval and analysis of mammographic masses," *Journal of Electronic Imaging*, vol. 14, no. 2, 2005, article no. 023016.

[7] Alberta Cancer Board, "Screen Test: Alberta Program for the Early Detection of Breast Cancer," 1999/01 Biennial Report, Edmonton, Alberta, Canada, 2001.

[8] K. Pelckmans, J. A. K. Suykens, T. V. Gestel, J. D. Brabanter, L. Lukas, B. Hamers, B. D. Moor, and J. Vandewalle, *LS-SVMlab1.5: Least squares support vector machines*, Katholieke University Leuven, Belgium, 2003.