Neurocomputing 72 (2009) 3296–3302

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Automatic detection of breast cancers in mammograms using structured support vector machines[☆]

Defeng Wang[a,*], Lin Shi[b], Pheng Ann Heng[a]

[a] Department of Computer Science and Engineering, The Chinese University of Hong Kong, N.T., Hong Kong
[b] Department of Diagnostic Radiology and Organ Imaging, The Chinese University of Hong Kong, N.T., Hong Kong

ARTICLE INFO

Article history:
Received 17 July 2008
Accepted 10 February 2009
Communicated by S. Hu
Available online 8 April 2009

Keywords:
Breast cancer detection
Computer-assisted diagnosis
Support vector machine
Structured support vector machine

ABSTRACT

Breast cancer is one of the most common cancers diagnosed in women. Large margin classifiers like the support vector machine (SVM) have been reported effective in computer-assisted diagnosis systems for breast cancers. However, since the separating hyperplane determination exclusively relies on support vectors, the SVM is essentially a *local* classifier and its performance can be further improved. In this work, we introduce a *structured* SVM model to determine if each mammographic region is normal or cancerous by considering the cluster structures in the training set. The optimization problem in this new model can be solved efficiently by being formulated as one second order cone programming problem. Experimental evaluation is performed on the Digital Database for Screening Mammography (DDSM) dataset. Various types of features, including curvilinear features, texture features, Gabor features, and multi-resolution features, are extracted from the sample images. We then select the salient features using the recursive feature elimination algorithm. The structured SVM achieves better detection performance compared with a well-tested SVM classifier in terms of the area under the ROC curve.

1. Introduction

Breast cancer is a very commonly diagnosed cancerous abnormality in that one third of all cancers detected among women are related to breast. Current technologies are very effective to treat early-stage breast cancers, which makes the early detection a crucial task [1]. Before clinical symptoms appear, screening mammography is still the most effective tool to catch early signs of cancerous abnormalities [2]. To improve the accuracy and efficiency of digital mammogram interpretation, computer-assisted diagnosis (CAD) [3] systems have been developed.

The objective of a typical breast cancer CAD system is to detect and evaluate various cancerous mammograms automatically. Fig. 1 illustrates the underlying principle of a CAD system. In this framework, the region of interest (ROI) is first selected from the mammogram as a sample image. Then the sample image is preprocessed by noise reduction and image enhancement [4]. A large number of features, such as texture features, multi-resolution features, and shape features, are obtained via feature extraction algorithms. To improve the classification efficiency, the redundant features are removed using a feature selection method.

There are three categories of abnormalities of interest in CAD systems, i.e., circumscribed masses [5], micro-calcification clusters [6], and spiculated or stellate lesions [7]. Example mammographic appearance in each category is illustrated in Fig. 2.

Founded upon Vapnik's statistical learning theory, the support vector machine (SVM) [8] has played an important role in many applications, including CAD based on medical images. Previous work, such as [9], has explored the use of SVM for detection of micro-calcification regions in digital mammograms. Even though SVM is a well-performed classifier, its performance is still limited because the data structure information is underutilized in the determination of the separating hyperplane. In this study, the detection of the three typical types of mammographic abnormalities (cf. Fig. 2) is formulated as a supervised learning problem. We propose the use of structured SVM (s-SVM) to detect breast cancers in digital mammograms. Fig. 3 shows an illustrative example of the intuition of proposing the s-SVM model. The square samples in the upper cluster tend to spread towards the opposite class, while the remaining ones scatter in the perpendicular direction. The standard SVM calculates the decision plane relying exclusively on the support vectors denoted by the light colored points, which results in an unbiased boundary separating the two classes. However, the structured SVM is designed to yield a decision plane leaving larger

[☆] This work was supported by the Hong Kong Research Grant Council under Grant CUHK 4453/06M.

[*] Corresponding author.
E-mail address: dfwang@cse.cuhk.edu.hk (D. Wang).

0925-2312/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2009.02.015

room for the upper cluster in the square class by integrating the cluster covariance information in the training process. Therefore, s-SVM is more likely to correctly classify the unseen patterns and thus potentially possesses good generalization ability.

2. Structured support vector machine

2.1. Data structure detection

To investigate the structure of a given dataset, the patterns are clustered hierarchically in the input space for the linear s-SVM and in the kernel space for the nonlinear s-SVM. The agglomerative hierarchical clustering (AHC) algorithm [10] can be described as follows.

```
Initialize each point as a cluster
Calculate the distance between every two clusters
While more than one cluster remains
  Find the closest pair of clusters
  Merge the two clusters
  Update the distance between each pair of clusters
End
```

The output of this algorithm is a tree structure known as the dendrogram [11], whose topology is also a representation of the clustering process. Therefore, by cutting this dendrogram at different levels, one can achieve diverse clustering results. Various hierarchical clustering approaches [10] differ in the method of finding the closest pair of clusters. We use the Ward's linkage clustering [12] in this study for the reason that clusters derived from this method are compact and spherical [13], which provides a meaningful basis for the calculation of covariance matrices. If S and T are two clusters with means $\bar{S}$ and $\bar{T}$, respectively, the Ward's linkage $W(S,T)$ between clusters S and T can be calculated as

$$W(S,T) = \frac{|S| \times |T| \times \|\bar{S} - \bar{T}\|^2}{|S| + |T|}. \tag{1}$$

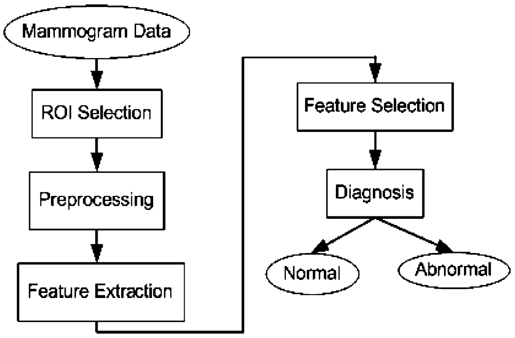

Fig. 1. Schematic block diagram of the system for breast cancer detection.

In the high-dimensional, implicit kernel space, the hierarchical clustering is still applicable. The Ward's linkage between $\Phi(\mathbf{x}_i)$ and $\Phi(\mathbf{x}_j)$, i.e., the images of patterns $\mathbf{x}_i$ and $\mathbf{x}_j$, can be calculated by

$$W(\Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j)) = \tfrac{1}{2}[k(\mathbf{x}_i,\mathbf{x}_i) + k(\mathbf{x}_j,\mathbf{x}_j) - 2k(\mathbf{x}_i,\mathbf{x}_j)],$$

where $k(\mathbf{x}_i,\mathbf{x}_j) := \Phi(\mathbf{x}_i)\cdot\Phi(\mathbf{x}_j)$ is a kernel function [8]. During hierarchical clustering, the Ward's linkage between clusters to be merged increases as the number of clusters decreases. A curve, namely the merge distance curve, is drawn to represent this process. The dendrogram can be cut when given the number of clusters, which can be determined by finding the knee point [14], i.e., the point of largest curvature on the merge distance curve.

2.2. second order cone programming (SOCP) formulation of s-SVM

Given a training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell}$ with input data $\mathbf{x}_i \in \mathbb{R}^n$ and class labels $y_i \in \{+1,-1\}$, the s-SVM calculates a decision hyperplane, i.e., $\mathbf{w}\cdot\mathbf{x} + b = 0$, to separate the two classes as robustly as possible. For the purpose of deriving structured SVM, we suppose there are C_P clusters in class P and C_N clusters in class N, i.e., $P = P_1 \cup \cdots \cup P_i \cup \cdots \cup P_{C_P}$ and $N = N_1 \cup \cdots \cup N_j \cup \cdots \cup N_{C_N}$. Since the clusters derived by the Ward's linkage AHC are compact and spherical [13], we assume each cluster has a Gaussian distribution, i.e., $P_i \sim \mathcal{N}\ (\boldsymbol{\mu}_{P_i}, \boldsymbol{\Sigma}_{P_i})$, $i = 1,\ldots,C_P$, and $N_j \sim \mathcal{N}\ (\boldsymbol{\mu}_{N_j}, \boldsymbol{\Sigma}_{N_j})$,

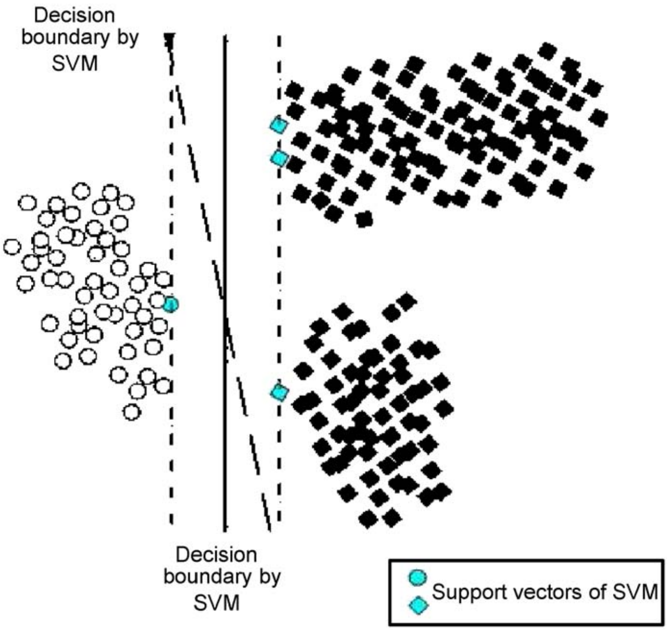

Fig. 3. The decision boundaries calculated by the SVM and the s-SVM. The s-SVM leaves more space for the upper cluster of the class represented by dark squares, since that cluster has stronger distribution tendency towards the opposite class.

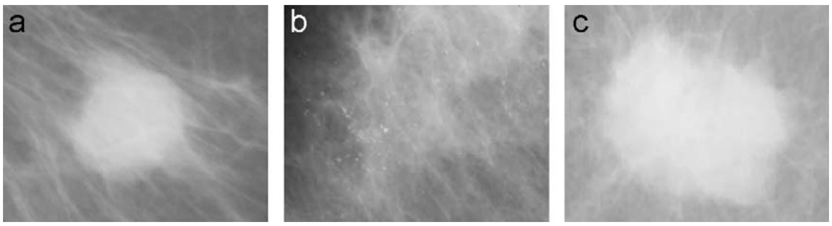

Fig. 2. Typical breast abnormalities in mammograms. (a) Circumscribed mass; (b) micro-calcification; (c) spiculated or stellate lesion.

$j=1,\dots,C_N$. The linear s-SVM can generally be formulated as

$$\begin{aligned} \min \quad & \sum_{\ell=1}^{|P|+|N|} \xi_\ell \\ \text{s.t.} \quad & \mathbf{Pr}\{(\mathbf{w}^T\mathbf{x}_\ell + b) \geqslant 1 - \xi_\ell\} \geqslant k_{P_i}, \quad \mathbf{x}_\ell \in P_i, \\ & \mathbf{Pr}\{-(\mathbf{w}^T\mathbf{x}_\ell + b) \geqslant 1 - \xi_\ell\} \geqslant k_{N_j}, \quad \mathbf{x}_\ell \in N_j, \\ & \|\mathbf{w}\| \leqslant \gamma, \\ & \xi_\ell \geqslant 0. \end{aligned} \tag{2}$$

Here we can find that the linear classification constraints in SVM optimization problem [8] are replaced by probabilistic ones. These probabilistic thresholds are determined proportional to the cluster size, i.e., $k_{P_i} = |P_i|/(|P|+|N|)$, and $k_{N_j} = |N_j|/(|P|+|N|)$. The probabilistic constraints in (2) can be restated as deterministic ones by utilizing the cluster distribution information [15]. Take the first probabilistic constraint for example, we define $z_\ell := -\mathbf{w}\cdot\mathbf{x}_\ell$, $\mathbf{x}_\ell \in P_i$. z_ℓ is a normal random variable with mean $\bar{z_\ell}$ and variance $\sigma_{z_\ell}^2 = \mathbf{w}^T\mathbf{\Sigma}_{P_i}\mathbf{w}$. Therefore $(z_\ell - \bar{z_\ell})/\sigma_{z_\ell} \sim \mathcal{N}(0,1)$. The first constraint then becomes

$$\mathbf{Pr}\left\{\frac{z_\ell - \bar{z_\ell}}{\sigma_{z_\ell}} \leqslant \frac{b - 1 + \xi_\ell - \bar{z_\ell}}{\sigma_{z_\ell}}\right\} \geqslant k_{P_i} \tag{3}$$

and we can compute the left-hand side of (3) by evaluating the cumulative distribution function for a standard normal Gaussian distribution

$$\psi(u) = \mathbf{Pr}\{\mathcal{N}(0,1) \leqslant u\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp(-s^2/2)\, ds.$$

Consequently, (3) is equivalent to the condition

$$\psi\left(\frac{b - 1 + \xi_\ell - \bar{z_\ell}}{\sigma_{z_\ell}}\right) \geqslant k_{P_i},$$

where $\psi(u)$ is monotonic increasing and invertible. As matrices $\mathbf{\Sigma}_{P_i}$ and $\mathbf{\Sigma}_{N_j}$ are positive semi-definite, there exist matrices $\mathbf{\Sigma}_{P_i}^{1/2}$ and $\mathbf{\Sigma}_{N_j}^{1/2}$ such that

$$\mathbf{\Sigma}_{P_i} = \mathbf{\Sigma}_{P_i}^{1/2^T}\mathbf{\Sigma}_{P_i}^{1/2}, \quad \mathbf{\Sigma}_{N_j} = \mathbf{\Sigma}_{N_j}^{1/2^T}\mathbf{\Sigma}_{N_j}^{1/2}. \tag{4}$$

Therefore, we have the following optimization problem:

$$\begin{aligned} \min \quad & \sum_{\ell=1}^{|P|+|N|} \xi_\ell \\ \text{s.t.} \quad & (\mathbf{w}^T\mathbf{x}_\ell + b) \geqslant 1 - \xi_\ell + \psi_{P_i}\|\mathbf{\Sigma}_{P_i}^{1/2}\mathbf{w}\|, \quad \mathbf{x}_\ell \in P_i, \\ & -(\mathbf{w}^T\mathbf{x}_\ell + b) \geqslant 1 - \xi_\ell + \psi_{N_j}\|\mathbf{\Sigma}_{N_j}^{1/2}\mathbf{w}\|, \quad \mathbf{x}_\ell \in N_j, \\ & \|\mathbf{w}\| \leqslant \gamma, \\ & \xi_\ell \geqslant 0, \end{aligned} \tag{5}$$

where $\psi_{P_i} := \psi^{-1}(k_{P_i})$ and $\psi_{N_j} := \psi^{-1}(k_{N_j})$. The optimization problem (5) is an instance of the second order cone programming. Minimizing a linear objective over second order cone (SOC) and linear constraints is known as an SOCP problem [16]. Recent advances in interior-point methods for the convex nonlinear optimization [17] have made such problems feasible. As a special case of the convex nonlinear optimization, SOCP problems have gained much attention recently, and can be handled efficiently by the existing software such as SeDuMi [18]. The total complexity of building the constraint matrix in the SOCP problem and solving the SOCP problem using interior-point method is $O((|P|+|N|)\cdot n^3)$ [16].

2.3. Connection with SVM

In s-SVM, we consider the data structure information in the hyperplane determination. Specifically, we assign a probabilistic value to each linear classification constraint based on the detected structures. If we ignore the data structure information and assume the constraints are defined with certainty as in the standard SVM model, the optimization problem (2) degenerates to

$$\begin{aligned} \min \quad & \sum_{\ell=1}^{|P|+|N|} \xi_\ell \\ \text{s.t.} \quad & (\mathbf{w}\cdot\mathbf{x}_\ell + b) \geqslant 1 - \xi_\ell, \quad \mathbf{x}_\ell \in P_i, \\ & -(\mathbf{w}\cdot\mathbf{x}_\ell + b) \geqslant 1 - \xi_\ell, \quad \mathbf{x}_\ell \in N_j, \\ & \|\mathbf{w}\| \leqslant \gamma, \\ & \xi_\ell \geqslant 0. \end{aligned} \tag{6}$$

For a proper choice of regularization parameter C in SVM [8] and the weight constraint parameter γ in s-SVM, the optimization problem (6) is equivalent to the one involved in SVM [19]. This means the SVM is a special case of the s-SVM.

2.4. Kernelization

For some problems, improved classification can be achieved using nonlinear s-SVM. According to Cover's pattern separability theory, patterns linearly nonseparable in the input space may be transformed into a kernel space to make them linearly separable, as long as the transformation is nonlinear and the dimensionality of the kernel space is high enough [20]. This nonlinear transformation can be achieved using Mercer kernels [8,21]. The basic idea of nonlinear s-SVM is to map data vectors from the input space to a high-dimensional feature space using a nonlinear mapping Φ, and then detect data structures via kernelized AHC before proceeding with pattern classification using linear s-SVM. However, the nonlinear mapping Φ is performed by employing kernel functions $k(\mathbf{x}_i,\mathbf{x})$, which obeys Mercer's conditions [8]. Thus the optimization problem of s-SVM in the kernel space can generally be formulated as follows:

$$\begin{aligned} \min \quad & \sum_{\ell=1}^{|P|+|N|} \xi_\ell \\ \text{s.t.} \quad & (\mathbf{w}^T\Phi(\mathbf{x}_\ell) + b) \geqslant 1 - \xi_\ell + \psi_{P_i}\|\mathbf{\Sigma}_{P_i}^{\Phi\,1/2}\mathbf{w}\|, \quad \mathbf{x}_\ell \in P_i, \\ & -(\mathbf{w}^T\Phi(\mathbf{x}_\ell) + b) \geqslant 1 - \xi_\ell + \psi_{N_j}\|\mathbf{\Sigma}_{N_j}^{\Phi\,1/2}\mathbf{w}\|, \quad \mathbf{x}_\ell \in N_j, \\ & \|\mathbf{w}\| \leqslant \gamma', \\ & \xi_\ell \geqslant 0. \end{aligned}$$

The above optimization problem is not solvable unless it is represented in the kernel form $k(\mathbf{x}_i,\mathbf{x}_j) = \Phi(\mathbf{x}_i)^T\Phi(\mathbf{x}_j)$, i.e., a dot product of maps of samples.

Corollary 1. *If the estimates of mean and covariance matrix of cluster S in the kernel space are, respectively,*

$$\boldsymbol{\mu}_S^\Phi = \frac{1}{|S|}\sum_{\mathbf{x}\in S}\Phi(\mathbf{x}) \tag{7}$$

and

$$\mathbf{\Sigma}_S^\Phi = \frac{1}{|S|}\sum_{\mathbf{x}\in S}(\Phi(\mathbf{x}) - \boldsymbol{\mu}_S^\Phi)(\Phi(\mathbf{x}) - \boldsymbol{\mu}_S^\Phi)^T, \tag{8}$$

the optimal $\mathbf{w}$ *lies in the space spanned by the training data maps.*

According to Corollary 1, we can write $\mathbf{w}$ as

$$\mathbf{w} = \sum_{\ell=1}^{|P|+|N|} \alpha_\ell \Phi(\mathbf{x}_\ell), \tag{9}$$

where $\alpha_\ell \in \mathbb{R}$ are coefficients. By using $\mathbf{w}$ expressed in terms of data images, we can obtain the kernel form of the optimization problem

$$
\begin{aligned}
\min \quad & \sum_{\ell=1}^{|P|+|N|} \xi_\ell \\
\text{s.t.} \quad & (\mathbf{K}_\ell \boldsymbol{\alpha} + b) \geqslant 1 - \xi_\ell + \psi_{P_i} \|\tilde{\mathbf{K}}_{P_i} \boldsymbol{\alpha}\|, \quad \mathbf{x}_\ell \in P_i, \\
& -(\mathbf{K}_\ell \boldsymbol{\alpha} + b) \geqslant 1 - \xi_\ell + \psi_{N_j} \|\tilde{\mathbf{K}}_{N_j} \boldsymbol{\alpha}\|, \quad \mathbf{x}_\ell \in N_j, \\
& \|\boldsymbol{\alpha}\| \leqslant \gamma, \\
& \xi_\ell \geqslant 0.
\end{aligned} \tag{10}
$$

$\mathbf{K}_\ell$ represents the ℓth row in the kernel Gram matrix $\mathbf{K}$, in which the elements satisfy $\mathbf{K}(i,j) = k(\mathbf{x}_i, \mathbf{x}_j)$, $i,j = 1, \ldots, |P| + |N|$. $\tilde{\mathbf{K}}_{P_i} = (1/\sqrt{|P_i|})(\mathbf{K}_{P_i} - \mathbf{e}_{|P_i|} \cdot \mathbf{v}_{P_i}^T)$, where $\mathbf{e}_{|P_i|}$ is an all-one column vector with length $|P_i|$. $\mathbf{K}_{P_i}$ is the kernel matrix between the cluster P_i and all the training patterns, that is $\mathbf{K}_{P_i}(s,j) = k(\mathbf{x}_s, \mathbf{x}_j)$, $s = 1, \ldots, |P_i|$, $j = 1, \ldots, |P| + |N|$. $\mathbf{v}_{P_i}$ is the mean vector of matrix $\mathbf{K}_{P_i}$ and $\mathbf{v}_{P_i}(j) = \sum_{\mathbf{x}_s \in P_i} k(\mathbf{x}_s, \mathbf{x}_j)/|P_i|$, $j = 1, \ldots, |P| + |N|$. $\tilde{\mathbf{K}}_{N_j}$ is calculated similarly to $\tilde{\mathbf{K}}_{P_i}$.

One can easily identify that this optimization problem is entirely expressed in terms of inner products between data images only, which makes the kernelized s-SVM solvable. The nonlinear constraints are exactly in the form of a second order cone programming problem [16]. The optimal separating hyperplane can be determined by solving the SOCP problem (10). Normally only a proportion of data points with coefficients α_ℓ are not zero, which are called the *support vectors* of the s-SVM. A test point $\mathbf{x}$ is discriminated as positive or negative by the following decision function:

$$
f(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b \right). \tag{11}
$$

The generalization error bound of s-SVM can be calculated according to Theorem 1 (see Appendix A for proof). Note that the generalization error bound subtracted from 1 gives the generalization bound.

Theorem 1. *Given a decision hyperplane calculated by s-SVM that separates the positive class P from the negative class N in the training set, and reliable estimates of means* $\boldsymbol{\mu}_{P_i}$ *(or* $\boldsymbol{\mu}_{N_j}$*) and covariance matrices* $\boldsymbol{\Sigma}_{P_i}$ *(or* $\boldsymbol{\Sigma}_{N_j}$*) for cluster* P_i *(or* N_j*),* $i = 1, \ldots, C_P$ *and* $j = 1, \ldots, C_N$*, the generalization error of s-SVM is bounded by*

$$
\sum_{i=1}^{C_P} \left(\frac{|P_i|}{|P| + |N|} \frac{1}{1 + d_{P_i}^2} \right) + \sum_{j=1}^{C_N} \left(\frac{|N_j|}{|P| + |N|} \frac{1}{1 + d_{N_j}^2} \right),
$$

where d_{P_i} *and* d_{N_j} *are the minimum Mahalanobis distances from the opposite half space to the cluster center* $\boldsymbol{\mu}_{P_i}$ *and* $\boldsymbol{\mu}_{N_j}$*.*

Actually, for nonlinear s-SVM, the minimum squared distance $d_{P_i}^2$ can be obtained by

$$
d_{P_i}^2 = \frac{\left(\dfrac{1}{|P_i|} \mathbf{e}_{P_i}^T \mathbf{K}_{P_i} \boldsymbol{\alpha} + b \right)_+}{\boldsymbol{\alpha}^T \tilde{\mathbf{K}}_{P_i}^T \tilde{\mathbf{K}}_{P_i} \boldsymbol{\alpha}}
$$

and $d_{N_j}^2$ by

$$
d_{N_j}^2 = \frac{\left(-\dfrac{1}{|N_j|} \mathbf{e}_{N_j}^T \mathbf{K}_{N_j} \boldsymbol{\alpha} - b \right)_+}{\boldsymbol{\alpha}^T \tilde{\mathbf{K}}_{N_j}^T \tilde{\mathbf{K}}_{N_j} \boldsymbol{\alpha}},
$$

where $(z)_+ = \max(z, 0)$.

Table 1
The training and test subsets each with normal and abnormal mammographic sample images.

	# Normal regions	Abnormal/cancerous regions		
		# Mass	# Calcification	# Spiculation
Training	150	26	28	27
Test	150	27	28	28
Total	300	53	56	55

3. Experimental validation

3.1. Description of the dataset

We use the benchmark dataset for testing mammography CAD algorithms, i.e., the Digital Database for Screening Mammography (DDSM) [22] to evaluate the proposed s-SVM model and make comparison with the standard SVM. DDSM is a publicly available database distributed by University of South Florida. The normal and abnormal regions each with 512 × 512 pixels were cropped from mammograms in DDSM. As indicated in [23], the sample image consisting of 512 × 512 pixels is large enough to contain commonly seen cancerous regions and to extract multi-resolution features. For the abnormal images, this 512 × 512 sample is centered in the center of the abnormal region. The normal sample images were semi-manually cropped from normal mammograms with various density types. Now the whole dataset contains 164 abnormal (or cancerous) and 300 normal sample images (see Table 1). In the abnormal subset, there are 53 masses, 55 micro-calcifications, and 56 spiculated lesions. We then partition the acquired dataset randomly into training and test subsets with similar proportions of cancerous and normal samples (see Table 1).

3.2. Preprocessing

Each mammographic sample image is preprocessed prior to feature extraction. We first use the median filtering to reduce the overall noise caused by statistics of X-ray quantum absorption [24]. Then the filtered sample images are enhanced using a physics-based mammogram enhancement method introduced in [4], which models the X-ray physics of the imaging process.

3.3. Feature extraction

A set of features is extracted from each mammographic region (or sample image) in the dataset because they all have been reported to be useful in separating cancerous regions from normal ones. These features include curvilinear features, texture features, Gabor features, and multi-resolution features [23].

- Curvilinear features: For the normal sample, the curvilinear markings tend to radiate from the nipple toward the chest wall. However, for the abnormal sample, the curvilinear structure usually appears as random or partially absent. A line detection algorithm was used to extract the curvilinear structure and 18 curvilinear features are extracted.
- Gray level co-occurrence features: The texture patterns in the mammographic regions can be well-characterized by the gray level co-occurrence matrix (GLCM). Sixteen texture features were extracted from the isotropic GLCM.

Table 2
Experimental results of all four combinations of the two classifiers (SVM and s-SVM) and the two kernel functions (the polynomial kernel and the Gaussian kernel).

	SVM-polynomial	SVM-Gaussian	s-SVM-polynomial	s-SVM-Gaussian
Parameters	$K = 2$, $C = 100$	$\sigma = 2.5$, $C = 500$	$K = 2.5$, $\gamma = 1$	$\sigma = 3$, $\gamma = 10$
# Features	23	19	20	16
A_z	0.907	0.942	0.939	0.970
Accuracy (%)	84.1	87.5	87.3	91.4
Lower bound (%)	–	–	72.5	85.4

- Gabor features: Gabor filter is capable of simultaneously dealing within both spatial and frequency domains. We use a total of 16 Gabor filters with the combination of four orientations and four scales to construct a Gabor filterbank. So for each region, 32 Gabor features are extracted.
- Multi-resolution statistical features: We use the Quincunx wavelet transform to extract multi-resolution features. Statistical features are extracted from the first four decomposition images in the even levels, which have the spatial resolution of 256×256, 128×128, 64×64 and 32×32. For each of them, five statistical features are extracted and a total of 20 multi-resolution features are extracted for each sample image.

For each sample image with 512×512 pixels, we generate an 86-feature-vector set by combining the above four feature sets. These features, together with the labels (normal or abnormal), constitute a supervised learning problem.

3.4. Feature selection via recursive feature elimination (RFE)

Although the cropped sample image is dramatically smaller in size compared with the original mammogram, the current feature vector is still not effective enough and considerably redundant for classification. So that a feature selection method is required to select a small set of the most discriminatory features for classification. A popular wrapper feature selection method, recursive feature elimination [25], is used in our current feature selection process. The RFE feature selection method was proposed for SVM to solve a cancer classification problem. RFE for SVM performs feature selection by iteratively training an SVM with the current set of features and removing the feature with the smallest weight in the resulting hyperplane at a time. Ultimately, this algorithm results in a ranking list of features.

3.5. Results

Polynomial and Gaussian kernels were tried with both SVM and s-SVM. To determine the parameters in both classifiers, we apply a 10-fold cross-validation [26]. The parameters with the smallest generalization error are chosen. With the best parametric setting obtained, i.e., the kernel function and associated parameters, the final form of the decision functions from the two classifiers are yielded by retraining them using all the training samples and testing afterwards using the test samples. The redundant features are eliminated using the RFE algorithm [25] for both classifiers with the two types of kernels. For each classifier with a certain kernel function, we initiate training the classifier with the top feature in the ranking, and increase the number of features following that ranking until the test accuracy begin to drop. The number of features is chosen to be the one corresponding to the highest test accuracy.

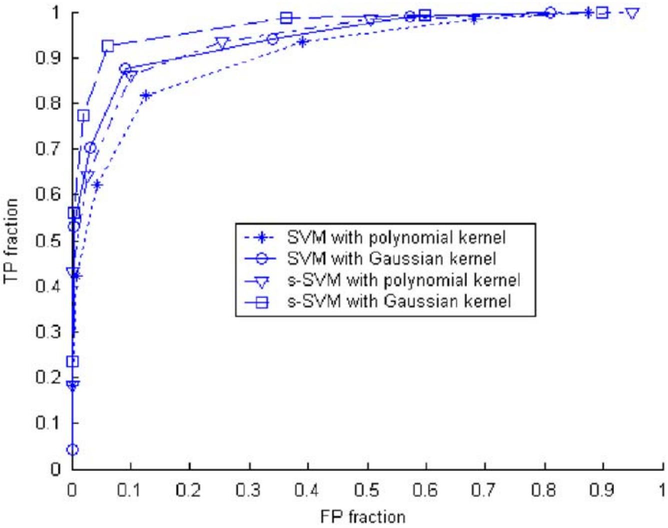

Fig. 4. The ROC curves of all four combinations of the two classifiers (SVM and s-SVM) and the two kernel functions (polynomial kernel and Gaussian kernel).

Now we evaluate SVM and s-SVM each with both Gaussian and polynomial kernels. The final parameter settings, the sizes of selected feature subsets, and the test accuracies are reported in Table 2. For s-SVM, we also list the calculated generalization lower bound, which equals the generalization error bound given in Theorem 1 subtracted from 1. The comparatively better performance of s-SVM validates our proposition that proper consideration of data structure does help to construct an accurate classifier. Moreover, in either classifier, using the Gaussian kernel does improve the performance compared to using the polynomial kernel. The optimal numbers of features selected using RFE are similar for all models. But for s-SVM and SVM using the same type of kernels, the former requires a smaller number of features compared to the latter.

The ROC curves for all combinations of the two classifiers and the two kernel types are plotted in Fig. 4 for an overall comparison. The areas under the ROC curves, i.e., A_z's in Table 2, also demonstrate that s-SVM does outperform SVM given the same type of kernels.

s-SVM outperforms SVM mainly because of its proper consideration of the data structure information. In order to demonstrate the existence of data structures in the kernel space, which is impractical to be directly displayed because of the infinite dimensionality, we choose to plot data images in the kernel space by kernel principal component analysis (KPCA) [27], i.e., projecting them onto the first three kernel principal components in the kernel space. The structures of the training dataset in both the polynomial kernel space and the RBF kernel

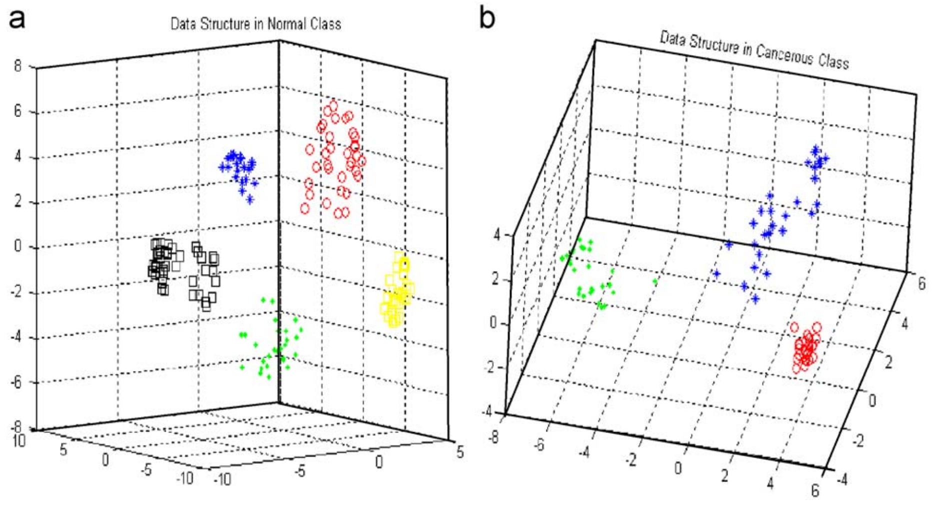

Fig. 5. The visualization of data structures in (a) the normal class; (b) the cancerous class by projecting the data images in the polynomial kernel space onto the three most principal kernel components.

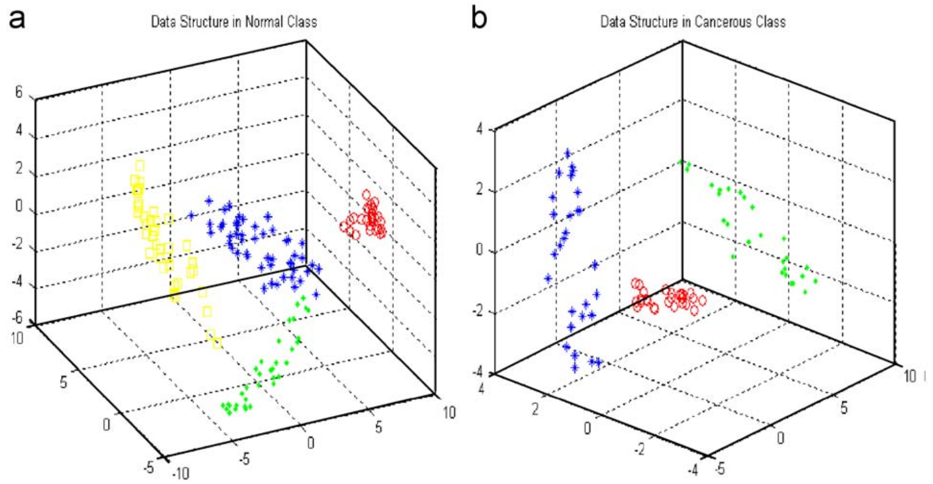

Fig. 6. The visualization of data structures in (a) the normal class; (b) the cancerous class by projecting the data images in the RBF kernel space onto the three most principal kernel components.

space are illustrated in Figs. 5 and 6, respectively. Note that the kernel parameter values are just the same as those determined in s-SVM training.

4. Conclusion

In this paper we proposed a structured large margin machine, i.e., s-SVM, by considering data structures in the training set. As the optimization problem in s-SVM can be formulated as only one second order cone programming problem, it can be solved efficiently. This new classifier is applied to determine if a sample image cropped from a digital mammogram is normal or cancerous based on features extracted by various methods and selected via the RFE algorithm. Experimental results show that s-SVM achieves generally better detection performance in comparison with the standard SVM.

In the future, we need to validate this classifier on more digital mammogram datasets to get concrete and systematic comparison with various large margin classifiers. Another important topic is the automatic determination of kernel type and the value for the kernel parameters.

Appendix A. Proof of Theorem 1

Proof. For the convenience of proof, we first recall the Marshall and Olkin Theorem [28,15] in Lemma 1.

Lemma 1 (*Marshall and Olkin Theorem*). *S is a convex set, and* **x** *is a random vector, then*

$$\sup_{\mathbf{x}\sim(\boldsymbol{\mu},\boldsymbol{\Sigma})} \mathbf{Pr}(\mathbf{x}\in S)=\frac{1}{1+d^2}\quad \textit{with } d^2=\inf_{\mathbf{x}\in S}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}),$$

where the supremum is taken over all distributions for **x** *with mean* $\boldsymbol{\mu}$ *and covariance matrix* $\boldsymbol{\Sigma}$.

In view of Lemma 1 and using $\mathscr{S}=\{\mathbf{w}\cdot\mathbf{x}+b\leqslant 0\}$, we obtain the upper bound on the probability that data in the positive cluster P_i are misclassified to the negative class

$$\sup_{\mathbf{x}\sim(\boldsymbol{\mu}_{P_i},\boldsymbol{\Sigma}_{P_i})} \mathbf{Pr}(\mathbf{w}\cdot\mathbf{x}+b\leqslant 0)=\frac{1}{1+d_{P_i}^2}\quad \text{with}$$

$$d_{P_i}^2=\inf_{\mathbf{w}\cdot\mathbf{x}+b\leqslant 0}(\mathbf{x}-\boldsymbol{\mu}_{P_i})^T\boldsymbol{\Sigma}_{P_i}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{P_i}).$$

Recently, in [15], a simple closed-form expression for the minimum distance d_{P_i} is derived as

$$d_{P_i}^2 = \inf_{\mathbf{w}\cdot\mathbf{x}+b\leqslant 0} (\mathbf{x}-\boldsymbol{\mu}_{P_i})^T \boldsymbol{\Sigma}_{P_i}^{-1} (\mathbf{x}-\boldsymbol{\mu}_{P_i}) = \frac{(\mathbf{w}\cdot\boldsymbol{\mu}_{P_i}+b)_+}{\mathbf{w}^T\boldsymbol{\Sigma}_{P_i}\mathbf{w}},$$

where $(z)_+ = \max(z, 0)$. Similarly, the minimum distance d_{N_j} is

$$d_{N_j}^2 = \inf_{\mathbf{w}\cdot\mathbf{x}+b\geqslant 0} (\mathbf{x}-\boldsymbol{\mu}_{N_j})^T \boldsymbol{\Sigma}_{N_j}^{-1} (\mathbf{x}-\boldsymbol{\mu}_{N_j}) = \frac{(-\mathbf{w}\cdot\boldsymbol{\mu}_{N_j}-b)_+}{\mathbf{w}\boldsymbol{\Sigma}_{N_j}\mathbf{w}}.$$

In nonlinear s-SVM, for the ith positive cluster P_i, using $\mathscr{S} = \{\mathbf{w}\cdot\Phi(\mathbf{x})+b\leqslant 0\}$, the upper bound on the probability that data in the positive cluster P_i are misclassified to the negative class is

$$\sup_{\mathbf{x}\sim(\boldsymbol{\mu}_{P_i}^{\Phi},\boldsymbol{\Sigma}_{P_i}^{\Phi})} \mathbf{Pr}(\mathbf{w}\cdot\Phi(\mathbf{x})+b\leqslant 0) = \frac{1}{1+d_{P_i}^{\Phi\,2}},$$

with

$$d_{P_i}^{\Phi\,2} = \inf_{\mathbf{w}\cdot\Phi(\mathbf{x})+b\leqslant 0} (\Phi(\mathbf{x})-\boldsymbol{\mu}_{P_i}^{\Phi})^T \boldsymbol{\Sigma}_{P_i}^{\Phi\,-1} (\Phi(\mathbf{x})-\boldsymbol{\mu}_{P_i}^{\Phi}). \tag{12}$$

Substituting (7)–(9) into (12), we derive

$$d_{P_i}^{\Phi\,2} = \frac{\left(\frac{1}{|P_i|}\mathbf{e}_{P_i}^T\mathbf{K}_{P_i}\boldsymbol{\alpha}+b\right)_+}{\boldsymbol{\alpha}^T\tilde{\mathbf{K}}_{P_i}^T\tilde{\mathbf{K}}_{P_i}\boldsymbol{\alpha}}.$$

Refer to Section 2.4 for the calculation of $\mathbf{K}_{P_i}$ and $\tilde{\mathbf{K}}_{P_i}$.

Similarly for the jth negative cluster, we have

$$d_{N_j}^{\Phi\,2} = \frac{\left(-\frac{1}{|N_j|}\mathbf{e}_{N_j}^T\mathbf{K}_{N_j}\boldsymbol{\alpha}-b\right)_+}{\boldsymbol{\alpha}^+\tilde{\mathbf{K}}_{N_j}^T\tilde{\mathbf{K}}_{N_j}\boldsymbol{\alpha}}.$$

Then the bounds on misclassification probability for positive and negative classes are $\sum_{i=1}^{C_P}((|P_i|/|P|)1/(1+d_{P_i}^{(\Phi)^2}))$ and $\sum_{j=1}^{C_N}((|N_j|/|N|)1/(1+d_{N_j}^{(\Phi)^2}))$, respectively. We assume the prior probability for any sample $\mathbf{x}$ belonging to the positive class to be $|P|/|P|+|N|$. Thus, the overall estimated error rate is

$$Ep_{err} \leqslant \sum_{i=1}^{C_P}\left(\frac{|P_i|}{|P|+|N|}\frac{1}{1+d_{P_i}^{(\Phi)^2}}\right) + \sum_{j=1}^{C_N}\left(\frac{|N_j|}{|P|+|N|}\frac{1}{1+d_{N_j}^{(\Phi)^2}}\right). \qquad \square$$

References

[1] T. O'Doherty, Review of effective image processing techniques of mammograms, Technical Report, Department of Information Technology, National University of Ireland, 1999.

[2] E.D. Andersen, A.D. Andersen, American Cancer Society: Cancer Facts and Figures 2005, American Cancer Society, Atlanta, 2005.

[3] L. Burhenne, S. Wood, C. D'Orsi, et al., Potential contribution of computer-aided detection to the sensitivity of screening mammography, Radiology 25 (2) (2000) 554–562.

[4] R. Highnam, M. Brady, Mammographic Image Analysis, Kluwer Academic Publishers, Dordrecht, 1999.

[5] T.O. Gulsrud, K. Engan, T. Hanstveit, Watershed segmentation of detected masses in digital mammograms, in: Proceedings of the 27th Annual Conference of the IEEE Engineering in Medicine and Biology, Shanghai, 2005.

[6] R.M. Nishikawa, M.L. Giger, K. Doi, C.J. Vyborny, R.A. Schimidt, Computer aided detection of clustered microcalcifications in digital mammograms, Med. Biol. Eng. Comput. 33 (1995) 174–178.

[7] S. Liu, C.F. Babbs, E.J. Delp, Multiresolution detection of spiculated lesions in digital mammograms, IEEE Trans. Image Process. 10 (6) (2001) 874–884.

[8] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1999.

[9] I. El-Naqa, Y. Yang, M.N. Wernick, N.P. Galatsanos, R.M. Nishikawa, A support vector machine approach for detection of microcalcifications, IEEE Trans. Med. Imaging 21 (12) (2002) 1552–1563.

[10] A.K. Jain, R. Dubes, Algorithms for Clustering Data, Prentice-Hall, New Jersey, 1988.

[11] S. Everitt, S. Landau, M. Leese, Cluster Analysis, Hodder Arnold, London, 2001.

[12] J.H. Ward, Hierarchical grouping to optimize an objective function, Journal of the American Statistical Association 58 (1963) 236–244.

[13] A. El-Hamdouchi, P. Willett, Comparison of hierarchic agglomerative clustering methods of document retrieval, The Computer Journal 32 (3) (1989) 220–227.

[14] S. Salvador, P. Chan, Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms, in: Proceedings of the 16th IEEE International Conference on Tools with AI, 2004, pp. 576–584.

[15] G.R.G. Lanckriet, L.E. Ghaoui, C. Bhattacharyya, M.I. Jordan, A robust minimax approach to classification, Journal of Machine Learning Research 3 (2002) 555–582.

[16] M. Lobo, L. Vandenberghe, S. Boyd, H. Lebret, Applications of second-order cone programming, Linear Algebra and its Applications 284 (1998) 193–228.

[17] Y. Nesterov, A. Nemirovskii, Interior-point polynomial algorithms in convex programming, in: Society for Industrial and Applied Mathematics (SIAM'94), Philadelphia, PA, USA, 1994.

[18] J. Sturm, Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones, Optimization Methods and Software 11 (1999) 625–653.

[19] C. Bhattacharyya, K.S. Pannagadatta, A. Smola, A second order cone programming formulation for classifying missing data, Advances in Neural Information Processing Systems (NIPS), vol. 17, 2004.

[20] S. Haykin, Neural Networks: A Comprehensive Foundation, Prentice-Hall International, New Jersey, 1999.

[21] B. Schölkopf, S. Mika, C. Burges, P. Knirsch, K.-R. Muller, G. Raetsch, A. Smola, Input space vs. feature space in kernel-based methods, IEEE Trans. Neural Networks 10 (5) (1999) 1000–1017.

[22] M. Heath, K. Bowyer, R. Kopans, D. Moore, J. Kegelmeyer, The digital database for screening mammography, in: 5th International Workshop on Digital Mammography, 2000, pp. 212–218.

[23] Y. Sun, C.F. Babbs, E.J. Delp, A comparison of feature selection methods for the detection of breast cancers in mammograms: adaptive sequential floating search vs. genetic algorithm, in: Proceedings of the 27th Annual Conference of the IEEE Engineering in Medicine and Biology, Shanghai, 2005.

[24] K. McLoughlin, P. Bones, N. Karssemeijer, Noise equalization for detection of microcalcification clusters in direct digital mammogram images, IEEE Trans. Med. Imaging 23 (2004) 313–320.

[25] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (2002) 389–422.

[26] K.R. Muller, S. Mika, K. Ratsch, G. Tsuda, B. Scholkopf, An introduction to kernel-based learning algorithms, IEEE Trans. Neural Networks 12 (2) (2001) 181–201.

[27] B. Schölkopf, A. Smola, K.-R. Muller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (1998) 1299–1319.

[28] A.W. Marshall, I. Olkin, Multivariate Chebyshev inequalities, Ann. Math. Statist. 31 (4) (1960) 1001–1014.