IMPROVING SVM CLASSIFIER WITH PRIOR KNOWLEDGE IN MICROCALCIFICATION DETECTION

Yan Yang, Juan Wang, and *Yongyi Yang*

Department of Electrical and Computer Engineering, Illinois Institute of Technology
3301 S. Dearborn Street, Chicago, IL 60616

ABSTRACT

This work aims to explore whether we can improve the accuracy of an SVM classifier for microcalcification (MC) detection by incorporating prior knowledge of MCs in mammograms. Based on the fact that MCs are inherently invariant to their spatial orientation in a mammogram, we consider two different techniques for incorporating rotation invariance into SVM, of which one is virtual support vector SVM (VSVM) and the other is tangent vector SVM (TV-SVM). The experiment results show that both techniques can improve the performance in discriminating MCs from the image background, and TV-SVM achieved the best performance. In particular, the sensitivity was 96.3% for TV-SVM, compared to 94.5% for SVM, when the false positive rate was at 0.5%.

Index Terms— Computer-aided diagnosis (CAD), support vector machine (SVM), virtual support vector SVM, tangent vector SVM.

1. INTRODUCTION

Clustered microcalcifications (MCs) can be an important early sign of breast cancer in women. MCs are tiny calcium deposits which appear as bright spots in mammograms (e.g, Fig. 1). They are found in 30%-50% of mammographically diagnosed cases [1]. While often seen, individual MCs can be difficult to detect due to their subtlety in appearance, variation in shape and size, and the surrounding breast tissue [2]. Accurate detection of MCs is a critical task in a computer-aided diagnosis (CAD) system, where the detected MCs are analyzed for subsequent classification into being benign or malignant.

Because of its importance, there has been significant interest in development of computerized methods for detection of clustered MCs in the literature. To name a few, these methods include the difference of Gaussian (DoG) detector [3], wavelet-based approaches [4]-[5], neural network [6], support vector machine (SVM) [7], and relevance vector machine [8]. Among them, the supervised learning methods have been the most popular.

In previous work [7] we developed an SVM approach for MC detection and demonstrated that it can outperform several representative detectors in the literature. In this approach, an SVM classifier was trained to discriminate image patterns of MCs from that of (non-MC) tissue regions of mammograms. These image patterns were obtained from a set of mammogram images used for training in which the MCs were pre-identified by experts. In Fig. 2 we show some examples of such image patterns of MCs and non-MCs randomly selected from a set of training images.

From Fig. 2 it can be seen that MCs indeed exhibit large degree of variation in appearance, including shape, size, and orientation. It is thus reasonable to expect that a robust MC detector should be insensitive to these variation factors in detection. Conceivably, such robustness could be obtainable provided that the detector is trained with a large set of samples that exhibit sufficient variations. However, this is impractical because the available training set is typically limited in size. Furthermore, the numerical complexity associated with classifier training typically increases more than linearly with the size of the training set.

In this work we investigate whether we can further improve the detection performance by exploiting spatial variation factors in a detector. In particular, we consider incorporating rotation invariance *a priori* into the SVM detector. This is because the nature of an MC object is independent of its spatial orientation in a mammogram. It is expected that incorporating this property into a detection algorithm could improve its robustness, thereby leading to improved detection accuracy.

In the literature there have been several techniques studied to incorporate invariance into SVM [9]-[11]. In this work, we consider two such techniques: virtual support vector SVM (VSVM) [9] and tangent vector SVM (TV-SVM) [11]. The basic idea of VSVM is to train the SVM with different variant versions (created virtually) of the training samples. In contrast, the TV-SVM is formulated such that its decision function is insensitive to variations in the input.

2. METHODS

In this study, MC detection is formulated as a two-class classification problem in which a supervised learning approach is applied. That is, for a given set of N training

This work was supported by NIH/NIBIB grant R01EB009905.

samples $\{(\mathbf{x}_i, y_i), i = 1, 2, \cdots, N\}$, where $\mathbf{x}_i$ is the feature vector of a training sample, and y_i is its known class label (1 for "MC present", and -1 for "MC absent"), we first train a classifier, and subsequently apply it to decide whether there is "MC present" or not at each location of a mammogram.

2.1 SVM for MC detection

Following our previous work [7], we consider SVM as the classifier in this study. To facilitate subsequent development, we briefly summarize the principle of this classifier below. For notional simplicity, we consider the case of a linear classifier function:

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b \tag{1}$$

where $\mathbf{w}$ and b are unknown and determined from the training dataset. This is achieved by the following formulation:

$$\min J(\mathbf{w}, \boldsymbol{\xi}) = \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{N} \xi_i \tag{2}$$

$$\text{s.t.} \quad y_i f(\mathbf{x}_i) \geq 1 - \xi_i, \ \xi_i \geq 0, i = 1, 2, \cdots, N.$$

In (2), the parameter C is used to control the trade-off between the model complexity (first term) and empirical risk (second term).

The above formulation can readily be extended to the case of a nonlinear classifier using the "kernel trick", where the input vector $\mathbf{x}$ is first mapped to a higher-dimensional space via a nonlinear mapping $\Phi(\mathbf{x})$, which is then classified by a linear classifier in this mapped space according to the criteria in (2).

With the kernel notion $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi^T(\mathbf{x}_i)\Phi(\mathbf{x}_j)$, the resulting SVM classifier function can be rewritten as:

$$f(\mathbf{x}_i) = \sum_{k=1}^{N_s} y_k \alpha_k K(\mathbf{x}_i, \mathbf{s}_k) + b \tag{3}$$

where $\mathbf{s}_k, k = 1, \cdots, N_s$, are the so-called support vectors, which correspond to those training samples that are either inside or on the decision margin of the classifier, which are determined during the training step.

For MC detection, we select the input vector $\mathbf{x}$ to be a small image window centered at a given location since individual MCs are well localized in a mammogram (Fig. 2). However, MCs exhibit a great deal of variation in their spatial orientation in mammograms. To exploit this variation in the detector, below we consider two techniques to incorporate rotation invariance *a priori* into the SVM classifier: one is the virtual support SVM [9] and the other is the tangent vector SVM [11].

2.2 Virtual support vector SVM

To incorporate rotation invariance into the SVM classifier, one straightforward approach would be to simply enlarge the training set by creating rotated versions of the available training samples (called "virtual examples"). While conceptually simple, this approach would also greatly increase the numerical complexity associated with SVM training, which is known to increase at least quadratically with the size of the training set [9].

Based on the property that the decision boundary of an SVM classifier is defined by only the support vectors, as seen from (3), and that the support vectors typically constitute a small portion of the training samples, a comprise is to generate rotated "virtual examples" only from the support vectors instead of all the training samples [9]. Such an approach is termed virtual support vector SVM (VSVM). Thus, a VSVM classifier is trained as follows:

1) train an SVM from the training set to obtain the support vectors;
2) generate "virtual examples" of the support vectors to form an new training set by combining the support vectors and their "virtual examples";
3) train a new SVM by the new training set.

In this approach, since only the support vectors contribute to the generation of "virtual examples", thus the invariance is not incorporated through the rest of the training samples. Conceptually, this may result in only a suboptimal solution. It would be more advantageous to take advantage of the invariance property of all the training samples in the classifier. The tangent vector SVM approach is used to address this issue [10]-[11].

2.3 Tangent vector SVM

For an input sample $\mathbf{x}_i$, it is desirable that the classifier output is invariant to a small perturbation (e.g., rotation) to $\mathbf{x}_i$. Specifically, let $L_t \mathbf{x}_i$ denote a perturbed version of $\mathbf{x}_i$ where L_t denotes the corresponding transformation (t is a parameter associated with the amount of transformation, e.g., rotation angle). Then for transformation invariance of $\mathbf{x}_i$, it is desired to have $f(L_t \mathbf{x}_i) = f(\mathbf{x}_i)$.

Consider the case of linear SVM in (1) at first. The above desired invariant property can be achieved when the vector $\mathbf{w}$ is orthogonal to the tangent vector $d\mathbf{x}_i$ associated with the transformation L_t to $\mathbf{x}_i$; the latter is defined as

$$d\mathbf{x}_i \equiv \lim_{t \to 0} \frac{1}{t} (L_t \mathbf{x}_i - \mathbf{x}_i). \tag{4}$$

To incorporate this property, the SVM objective function in (2) is modified as

$$J(\mathbf{w}, \boldsymbol{\xi}) = \frac{1}{2} \left(\gamma \| \mathbf{w} \|^2 + (1 - \gamma) \sum_{i=1}^{N} \left(\mathbf{w}^T d\mathbf{x}_i \right)^2 \right) + C \sum_{i=1}^{N} \xi_i \tag{5}$$

where the parameter γ (between 0 and 1) is introduced to control the trade-off between the normal SVM and invariance.

Similarly, by applying the kernel trick, we can extend the above tangent vector formulation to the case of a nonlinear classifier. For brevity, the details are omitted here and the interested reader is referred to [11]. A potential difficulty associated with this formulation is the increased computational complexity due to the high dimensionality of

the kernel space. To address this issue, the technique of kernel PCA is applied [11].

2.4 Performance evaluation

a) Mammogram dataset

To evaluate the performance of the different classifiers, we make use of a dataset collected by the Department of Radiology at the University of Chicago. It consists of a total of 200 different mammogram images, all containing multiple MCs. These images are of dimension 1024×1024 or 512×512 pixels, digitized with a spatial resolution of 0.1mm/pixel and 10-bit grayscale. The MCs in each mammogram were manually identified by a group of experienced radiologists. There are overall 5,211 MCs in these mammograms. These MCs were used as the ground truth in our experiments.

To suppress the inhomogeneity of the background in each mammogram, the high-pass filter [7] was first applied to remove the background. Afterward, a noise equalization procedure [12] was applied for equalizing the intensity-dependent noise in the image.

In our experiments, the dataset was randomly partitioned into two subsets, one with 50 mammograms for training and the other with 150 mammograms for testing.

b) Formation of training and testing samples

To characterize the image content of an MC object, we use a small image window centered at the location of the MC as in [7]. To facilitate the generation of rotated samples for VSVM and TV-SVM, we choose a circular window instead. The image pixels within this circular window are then used to form a vector as input to the SVM.

The MC samples are extracted from the known MCs in each mammogram image and used as "MC present" samples. Similarly, twice as many "MC absent" samples are randomly extracted from the background tissue area of the mammogram images. To evenly distribute the "MC present" examples across the different mammograms for training, no more than 40 such "MC present" samples are included for each image in the training set. For the mammograms in the testing set, all MCs are used. In summary, in our experiments a total of 956 MC samples were extracted for the training set, and 4,159 MC samples were extracted for the testing set.

To generate rotated virtual samples for VSVM, we applied incremental rotation by every 45 degrees for both extract "MC present" and "MC absent" samples in the training set.

For TV-SVM, the parameter γ in (5) was set as in [11]

$$\gamma = \frac{\alpha}{S + \alpha(1-S)}, \quad S = \sum_{i=1}^{N} \left\| d\mathbf{x}_i \right\|^2$$

where α was set to 0.9 in our experiments.

c) Kernel function

For the SVM, the following two kernel functions are used:

1) Polynomial kernel

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{y} + 1)^p \tag{6}$$

where p>0 is a constant that defines the kernel order.

2) Gaussian RBF

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\| \mathbf{x} - \mathbf{y} \|^2}{2\sigma^2} \right) \tag{7}$$

where σ>0 is a constant that defines the kernel width.

d) Testing and evaluation

To determine the optimal parameters of the classifier, we applied a four-fold cross validation procedure on the training set of image samples. From this procedure, the parametric setting with the smallest generalization error was chosen, and the classifier was retrained with all the training samples with this setting, and then applied to the testing dataset.

To evaluate the detection performance, we conduct a receiver operating characteristic (ROC) analysis, which is now routinely used for performance evaluation in classification tasks. An ROC curve is a plot of the classification sensitivity (i.e., true positive fraction) versus the specificity (i.e., false positive fraction) with the decision threshold continuously varied over its operating range.

3. RESULTS

3.1 SVM training and model selection

Fig. 3 shows the generalization error results obtained by the TV-SVM classifier with a Gaussian RBF kernel, where the generalization error is plotted versus the regularization parameter C for kernel width σ=25, 50 and 100. As can be seen, the generalization error is rather insensitive to these parameters, and the best error level is achieved when σ=25 and C=0.1. Similarly, generalization error curves were also obtained for both SVM and VSVM, but omitted here for brevity. In Table I, we summarize the optimal generalization error levels achieved by SVM, VSVM, and TV-SVM and their corresponding settings with Gaussian RBF kernel.

Similar results were also achieved with the polynomial kernel but not shown in favor of brevity. Below we will show results obtained with the Gaussian kernel.

3.2 Classification performance

In Fig. 4 we summarize the classification performance results achieved by SVM, VSVM and TV-SVM when applied to the test set, where the ROC curves are shown for FPF in the range of [0, 0.01]. Note that an FPF rate above 1% is considered to be too high to be useful in MC detection, because the overwhelming majority of pixels in a mammogram image are from the non-MC class.

From Fig. 4, the following can be observed: 1) for FPF higher than 0.3%, VSVM and TV-SVM are similar in detection performance, and both are higher in true detection rate than SVM; 2) for FPF lower than 0.3%, VSVM becomes similar to SVM, and TV-SVM is notably higher in true detection rate than both of them. These results indicate

that TV-SVM can be more effective for improving the detection accuracy through explicit incorporation of invariance into its decision function.

Finally, as a test of invariance of the trained VSVM and TV-SVM classifiers to input rotation, we randomly selected 1,500 MC samples from the test set. For each sample, its rotated versions were created by every 40 degrees and fed to the classifier; the standard deviation of the classifier output over these rotated versions was computed, and averaged over the 1,500 samples. The resulting standard deviation values are 0.0356, 0.0028, and 0.0028 for SVM, VSVM, and TV-SVM, respectively. Indeed, VSVM and TV-SVM are insensitive to the variation in orientation.

4. CONCLUSION

In this paper, we investigated the use of two techniques for incorporating orientation invariance *a priori* into an SVM classifier for MC detection in mammograms. The results show that this can be beneficial in improving the detection accuracy. Of the two techniques considered, the tangent vector SVM was demonstrated to be more effective for applications with low false-positive rate. In future studies, it would be interesting to further investigate whether incorporation of other invariance properties (e.g., shape or size) could yield further improvement.

5. REFERENCES

[1] American Cancer Society, "Cancer facts and figures", Atlanta, GA, 2009.

[2] M. Lanyi, "Diagnosis and differential diagnosis of breast calcifications," Berlin, Germany: Springer-Verlag, 1988.

[3] J. Dengler, S. Behrens, and J. F. Desaga, "Segmentation of microcalcifications in mammograms," *IEEE Trans. Med. Imag.,* vol. 12, pp. 634–642, Dec. 1993.

[4] R. N. Strickland and H. L. Hahn, "Wavelet transforms for detecting microcalcifications in mammograms," *IEEE Trans. Med. Imag.,* vol. 15, pp. 218–229, Apr. 1996.

[5] C. H. Chan and G. G. Lee, "On digital mammogram segmentation and microcalcification detection using multiresolution wavelet analysis," *Graphical models and image processing,* vol. 59, no. 5, pp. 349-364, Sep. 1997.

[6] S. Yu and L. Guan, "A CAD system for the automatic detection of clustered microcalcifications in digitized mammogram films," *IEEE Trans. Med. Imag.*, vol 19, no. 2, Feb. 2000.

[7] I. E1-Naqa, Y. Yang, M. N. Wernick, N. P. Galasanos, and R. M. Nishikawa, "A support vector machine approach for detection of microcalcifications," *IEEE Trans. Med. Imag.*, vol. 21, no. 12, Dec. 2002.

[8] L. Wei, Y. Yang, R. M. Nishikawa, M.N. Wernick, and A. Edwards, "Relevance vector machine for automatic detection of clustered microcalcifications" *IEEE Trans. Med. Imag.*, vol. 24, no.10, Oct. 2005.

[9] D. Decoste, "Training invariant support vector machines," *Machine Learning*, vol. 46, pp. 161-190, 2002.

[10] B. Schölkopf, P. Simard, A. Smola, and V. Vapnik, "Prior knowledge in support vector kernels," *Advances in Neural Information Processing Systems 10*, pp. 640-646, 1997.

[11] O. Chapelle, B. Schölkopf, "Incorporating Invariances in non-linear support vector machines." *Advances in Neural Information Processing Systems 10*, pp. 609-616, Dec. 2001.

[12] W. J. H. Veldkamp and N. Karssemeijer, "Normalization of local contrast in mammograms," *IEEE Trans. Med. Imag.,* vol. 19, no. 7, July 2000.

Table I Optimal settings of SVM, VSVM and TV-SVM.

Classifiers	SVM	VSVM	TV-SVM
Generalization Error	0.0132	0.0108	0.0087
Parameter σ	50	50	25
Parameter C	10	1	0.1

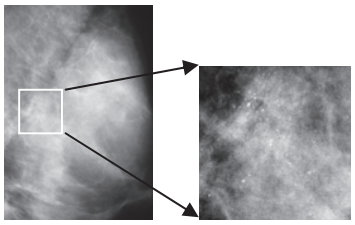

Fig. 1 A mammogram (left) and a magnified view (right).

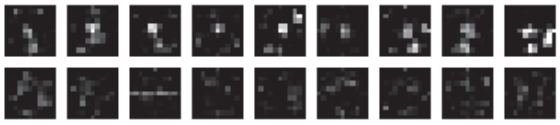

Fig. 2 MC (1st row) and non-MC (2nd row) samples.

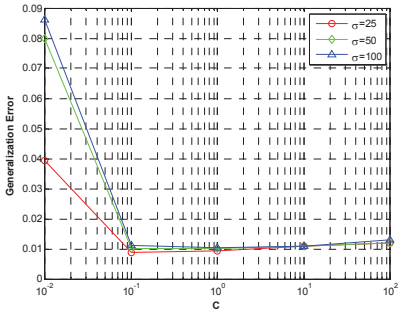

Fig. 3 Generalization error versus regularization parameter C for TV-SVM with Gaussian RBF kernel.

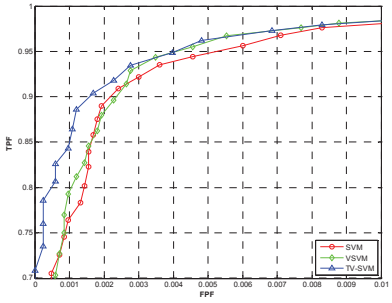

Fig. 4 ROC curves of SVM, VSVM, TV-SVM.