SVM Based Feature Screening Applied To Hierarchical Cervical Cancer Detection *

Jiayong Zhang, Yanxi Liu and Tong Zhao
The Robotics Institute, Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

We present a novel feature screening method by deriving relevance measures from the decision boundary of Support Vector Machine, which has several advantages over traditional screening methods based on Information Gain and Augmented Variance Ratio. The new algorithm is applied to a bottom-up approach to cervical cancer detection in multispectral PAP smear images that has been recently proposed by the authors. Comparative experiments show significant improvements on pixel-level classification accuracy using the new feature screening method.

1 Introduction

Finding abnormal cells in PAP smear images is a "needle in a haystack" type of problem, which is tedious, labor-intensive and error-prone. It is therefore desirable to have an automatic screening tool such that human experts are only called for when complicated and subtle cases arise. Most researches to date on automatic cervical screening try to extract morphometric/photometric features at the cellular level in accordance with "The Bethesda System" rules [7]. They usually depend on accurate segmentations between not only background and cells, but also cytoplasm and nucleus. However, various uncertainty factors make such segmentations rather difficult.

Recently, we have proposed a bottom-up approach to this problem via multi-level image classification without the requirement of accurate segmentation [8], an overall picture of which is given in Figure 1. Pixel-level analysis has been identified as the most important part of the system, where two critical issues exist: (1) what features should be extracted from multispectral images, and (2) how to remove irrelevant and/or redundant features from a pool of thousands of potential features to locate a feature subset that is well balanced between performance and compactness. For the first issue, we have identified a feasible feature space of about 4,000 dimensions that well captures local multispectral and texture information [9]. For the second issue, given that 4,000 dimensions is still intractable for traditional feature selection methods, we have employed in our previous work [8] two simple feature screening measures, i.e. *Information Gain* (IG) and *Augmented Variance Ratio* (AVR), to rule out irrelevant features. However, this method has an underlying limitation, that is, they are unable to model strong correlations between features since IG and AVR are evaluated independently for each feature. Thus it is possible to miss some combinations of discriminative but highly correlated features.

In this paper, we present a novel feature screening method by deriving relevance measures from the decision boundary of Support Vector Machine [2]. The proposed method has several advantages: 1) As the relevance measures are derived simultaneously for all dimensions, they do not have the "independence" problem of IG and AVR; 2) The maximum

*This research is supported in part by PA state grant ME#01-738 on "Cancer Informatics: From Molecules to Clinical Outcomes", an NIH National Cancer Institute Unconventional Innovation Program, award # N01-CO-07119, and in part by an NSF research grant #IIS-0099597.

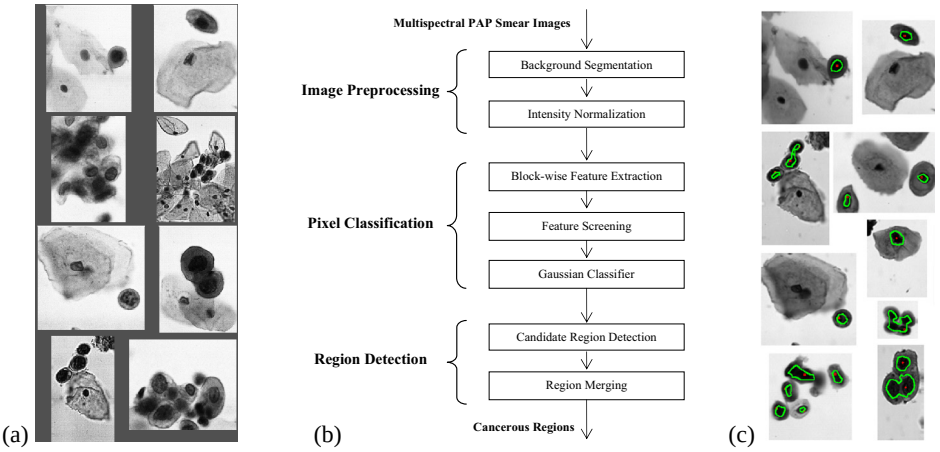

Figure 1: A bottom-up approach to cervical cancer detection recently proposed by the authors [8]. It took advantages of multispectral texture features without the requirement of accurate segmentation. The system was demonstrated on a multispectral PAP smear image database collected by a micro-interferometric spectral imaging setup at CMU [4], with wavelength ranging from 400 nm to 690 nm, evenly divided into 52 bands. (a) Sample (average-band) images from the database. (b) The hierarchical structure of the system. (c) Some detection results, with contours of detected cancerous regions overlapped on average-band images.

margin boundary provided by SVM has been proven to be optimal in a structural risk minimization sense, thus the new relevance measures better indicate the discriminative power of features; 3) As efficient routines for SVM training are available that can readily deal with huge number of features and samples, the proposed screening method does not sacrifice in computational cost. Comparative experiments with our original system show significant improvements on pixel-level classification accuracy by using the new screening method.

In addition, we have explored the features remained after the SVM based screening process by *sequential backward selection* (SBS), which leads to further reduction in subset sizes. Analysis of the selected feature subsets with respect to their extraction methods is provided in an attempt to get some insight into the interpretations of the selection results.

2 SVM Based Feature Screening

Given a set of features in a classification problem, a basic quesion in many learning tasks is: what is the best feature subset for classification purpose? Although many feature subset selection methods have been proposed [6, 1], few of them can be directly applied to domains with more than 100 dimensions. The huge feature dimension (near 4,000) and sample complexity (over 100,000) in our task make them computationally prohibitive. Alternatively, we present a new feature screening algorithm by deriving relevance measures from the decision boundary of Support Vector Machine. Features are ranked according to these measures, and then a subset is readily selected via some statistical significance test.

The decision function of a two-class problem derived by SVM can be written as

$$h(x) = w \cdot \Phi(x) + b = \sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b \tag{1}$$

where $x_i \in \mathbb{R}^d$ is the training sample, and $y_i \in \{\pm 1\}$ is the class label of x_i. A transformation $\Phi(\cdot)$ maps the data points x of the input space $\mathbb{R}^d$ into a higher dimensional feature space $\mathbb{R}^D$, $(D \geq d)$. The mapping is performed by a kernel function $K(\cdot,\cdot)$ which defines an inner product in $\mathbb{R}^D$. The parameters $\alpha_i \geq 0$ are optimized by finding the hyperplane in feature space with maximum distance to the closest image $\Phi(x_i)$ from the training set, which reduces to solving a linearly constrained convex quadratic program.

In the general case of nonlinear mapping Φ, SVM generates a nonlinear boundary $h(x) = 0$ in the input space. Given any two points $z_1, z_2 \in \mathbb{R}^d$ such that $h(z_1)\, h(z_2) < 0$, a surface point $s = \alpha z_1 + (1-\alpha) z_2, \alpha \in [0,1]$, can be found by solving the following equation with respect to α :

$$h(s) = h\left(\alpha z_1 + (1-\alpha) z_2\right) = 0 \tag{2}$$

The unit normal vector $N(s)$ at the boundary point s is then given by

$$N(s) = \nabla h(s) / \|\nabla h(s)\| \tag{3}$$

where $\nabla h(s) = \partial h(s)/\partial s = \sum_{i=1}^{n} \alpha_i y_i\, \partial K(s, x_i)/\partial s$. $N(s)$ identifies the orientation in the input space on which the projected training data are well separated locally around the neighborhood of s. Therefore, the orientation difference between $N(s)$ and any direction u can be used to measure the local discriminative relevance for that direction at s. Formally, we can measure this difference by $|u^T N(s)|$, or equivalently $u^T N(s) N(s)^T u$. To summarize all the local feature relevance information, we can compute the so-called *decision boundary scatter matrix* (DBSM) as

$$M = \int_{\mathcal{B}} N(s) N^T(s) p(s)\, ds \tag{4}$$

from which a global relevance measure for direction u can be computed as $u^T M u$. When sample-size is finite, M can be replaced by the sample estimate $\hat{M} = \sum_{i=1}^{l} \hat{N}(\hat{s}_i)\, \hat{N}(\hat{s}_i)^T / l$, where $\hat{s}_i$ are l points sampled from the estimated decision boundary. This global relevance measure can be readily extended to multi-category problems by repeating the procedure in either one-vs-all or pairwise mode. Now we can summarize the SVM based feature screening algorithm in Figure 2.

Several issues in the algorithm need some explanation. First, we prune those training samples far away from the decision boundary in locating the boundary points. This helps to reduce computational cost and suppress the negative influence of outliers. Second, we adopt the one-vs-all approach for solving Q-class problems with SVMs. Totally Q SVMs need to be trained, each of which separates a single class from all remaining classes. Third, the complexity of SVM-DBA can be controlled by several parameters including l, the number of boundary points to be sampled, and ϵ, the accuracy of the root to equation (2). Our experience seems to suggest that SVM-DBA is not very sensitive to the choice of these parameters. Finally, we have used p-degree polynomial kernels in our experiments.

It can be proven that a feature u is irrelevant if and only if $u^T M u$ equals zero. In theory we can exactly prune all irrelevant features via this screening method. However, inevitable errors in our estimation prevent us from doing so. A more practical reason is that, features' contribution to discrimination may be so unevenly distributed that the subset dimension can be significantly reduced while achieving *almost* the same accuracy. Therefore other model selection technique is required in order to decide an appropriate subset. This problem will not be discussed in this paper, but we want to point out that nested subsets generated by SVM based screening can easily facilitate such explorations.

Input: n sample pairs $\{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{k\}_{k=1}^{Q}$.

Output: d nested feature subsets $\mathcal{S}_1 \subset \mathcal{S}_2 \subset \ldots \subset \mathcal{S}_d$ such that $\dim(\mathcal{S}_m) = m$.

Algorithm:

S1 Repeat **S2** to **S5** for $k = 1$ to Q.

S2 Divide the n samples into two subsets $T^+ = \{x_i | y_i = k\}$ and $T^- = \{x_i | y_i \neq k\}$. Learn a SVM decision function $h(x)$ using T^+ and T^-.

S3 Sort the n samples in ascending order by the absolute function output values $|h(x_i)|$. Denote the subset consisting of the first r samples as T'.

S4 Select l pairs of points $\{(z_1^j, z_2^j)\}_{j=1}^{l}$ from T' randomly such that $h(z_1^j)\, h(z_2^j) < 0$. For each pair solve equation (2) to an accuracy of ϵ, and thus get l estimated boundary points $\{\hat{s}_j\}_{j=1}^{l}$.

S5 Compute the unit surface norm $\hat{N}(\hat{s}_j)$ at $\hat{s}_j$ according to equation (3), and estimate the decision boundary scatter matrix as $\hat{M}_k = \sum_{j=1}^{l} \hat{N}(\hat{s}_j)\hat{N}(\hat{s}_j)^T$.

S6 Compute $\hat{M} = \sum_{k=1}^{Q} \hat{M}_k / Q$, and denote its diagonal value as $\{\hat{\lambda}_j\}_{j=1}^{d}$.

S7 Sort feature directions $\{u_j\}_{j=1}^{d}$ descendingly by $\{\hat{\lambda}_i\}_{j=1}^{d}$. Let $\mathcal{S}_m = \{u_j^{sorted}\}_{j=1}^{m}$.

Figure 2: The proposed SVM based feature screening algorithm.

Guyon et al. [5] proposed a feature ranking scheme by linear SVMs. The basic idea is to use the magnitude of the weights of a linear discriminant classifier as an indicator of feature relevance. Our method can be considered as a nonlinear extension of this linear scheme. SVM boundary has also been used in locally adaptive metric techniques to improve k-NN performance [3]. Measures of local feature relevance are computed by the surface normal near the query, from which a local full-rank transformation is derived. Such local methods need to perform k-NN procedure multiple times in the original high-dimensional space. On the contrary, our method tries to globally characterize the discriminative information embedded in the SVM decision boundary. It generates global feature relevance measures, and thus is computationally much more efficient.

3 Experiments and Analysis

3.1 Pixel-level classification comparison

We used the same experimental setup as in [8] to investigate the effect on pixel-level classification by replacing IG and AVR feature screening with the proposed method. We started from 158,127 samples from 40 images (26,064 positive and 132,063 negative). In order to reduce the training complexity, a total of 29,487 samples were randomly selected (13,022 positive and 16,465 negative). SVM based feature screening method with p-degree polynimial kernels was applied to each of four types of wavelet features respectively. For each type of wavelet, images were randomly divided into training set (32 images) and test set (8 images) for a number of times. Each time we recorded the curve of *false positive rate* (FPR) on the test set versus subset dimension. Then we averaged these FPR curves, based on which a proper dimension m was hand selected. After that we collected all features ever appeared among the top m features in each image partition, and regarded them as the selected features for that wavelet type. Then we put together all the selected features for four types

of wavelets, and applied the SVM based feature screening algorithm. Again we did random partition of training and test images, and selected a proper dimension m' based on the average FPR curve. Various dimensions before and after feature screening are summarized in Table 1.

Finally we evaluated the selected 68 features on the original full sample set using the *modified quadratic discriminant function* (MQDF). The ROC curve is plotted in Figure 3 with the ROC curve of IG+AVR screening depicted for comparison. It is easy to observe that SVM based screening outperforms IG+AVR, especially when the *true positive rate* (TPR) is high.

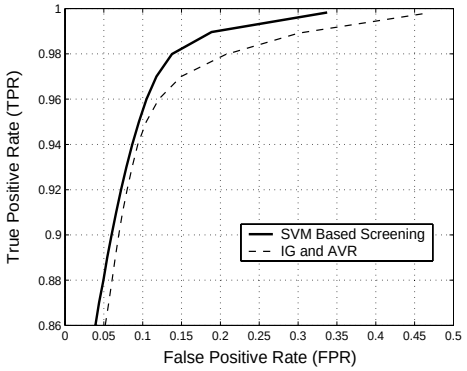

Figure 3: ROC comparison between SVM and IG+AVR based screenings.

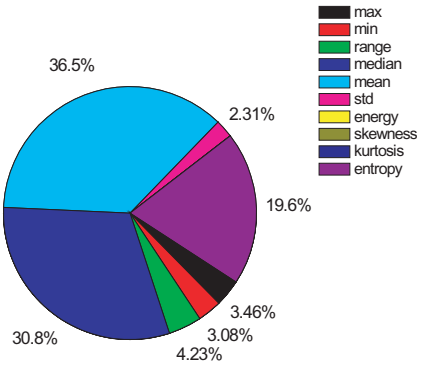

Figure 4: Pie plot of freq. of various statistics in SBS output.

3.2 Sequential backward selection

We applied SBS to the 68 features selected by SVM based screening to investigate their redundancy. 8-fold cross validation error of MQDF on the training set was chosen as the evaluation function in SBS. We averaged the test set FPR curves over 13 runs and depicted in Fig 5. It can be observed that feature dimension can be consistently reduced below 40 with little loss of accuracy. We analyzed those features that rank among the top 40 in any of the 13 runs with respect to their extraction methods. Figure 4, 6 and 7 show the frequencies of their appearances grouped in statistics type, wavelet type and multispectral band respectively. It can be observed that distributions of discriminative features are not uniform. How to interprete the selection results still deserves further study.

Table 1: Various dimensions before and after SVM based feature screening.

	DB2	DB16	Bio2.2	Gabor	Combined
Original	800	800	900	1200	3700
After Screening	48	42	52	30	68

4 Conclusion

In this paper, we presented a novel SVM-based feature screening method and applied it to multispectral Pap smear image classification for cervical cancer detection. Comparative

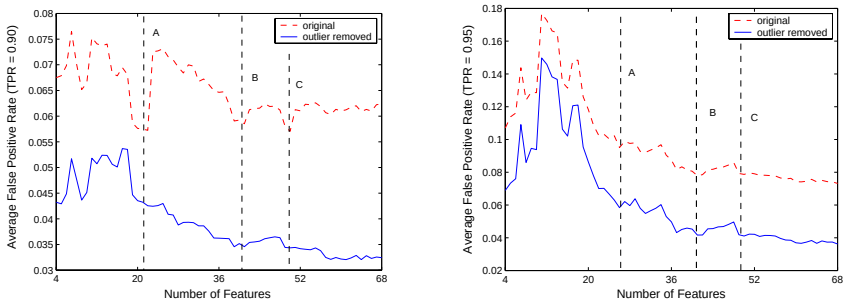

Figure 5: Average FPR curves with respect to subset dimensions in SBS. (Left) TPR = 0.90, (Right) TPR = 0.95.

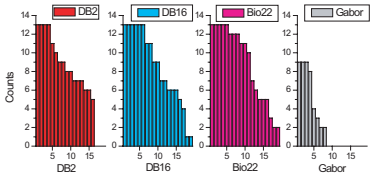

Figure 6: Freq. histograms of various wavelet types in SBS output.

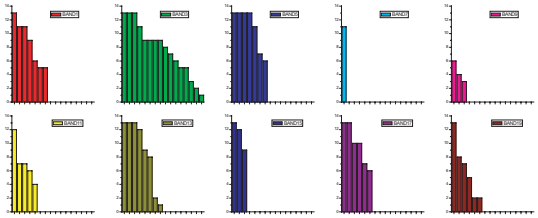

Figure 7: Freq. histograms of various bands in SBS output.

experiments show significant improvements on pixel-level classification accuracy using the new feature screening method. A much larger PAP smear image set and an even richer image feature space will be used to further validate our method.

References

[1] A. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271, 1997.

[2] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK, 2000.

[3] C. Domeniconi and D. Gunopulos. Adaptive nearest neighbor classification using support vector machines. In *Advances in Neural Information Processing Systems 14*. MIT Press, 2001.

[4] D. Farkas, B. Ballou, et al. Optical image acquisition, analysis and processing for biomedical applications. *Springer Lecture Notes in Computer Science*, 1311:663–671, 1997.

[5] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

[6] G. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *Proceedings of the International Conference on Machine Learning*, pages 121–129, 1994.

[7] R. Kurman and D. Solomon. *The Bethesda System for Reporting Cervical/Vaginal Cytologic Diagnoses*. Springer-Verlag, New York, 1994.

[8] Y. Liu, T. Zhao, and J. Zhang. Learning multispectral texture features for cervical cancer detection. In *Proceedings of 2002 IEEE International Symposium on Biomedical Imaging: Macro to Nano*, 2002.

[9] T. Zhao, J. Zhang, and Y. Liu. Does multispectral texture features really improve cervical cancer detection? In *International Conference on Diagnostic Imaging and Analysis*, 2002.