SVM Optimization for Hyperspectral Colon Tissue Cell Classification

Kashif Rajpoot[1] **Nasir Rajpoot**[2]

[1] *Faculty of Computer Science & Engineering, GIK Institute, Topi, PAKISTAN*
[2] *Department of Computer Science, University of Warwick, Coventry, UK*

emails: kmr@giki.edu.pk, nasir@dcs.warwick.ac.uk

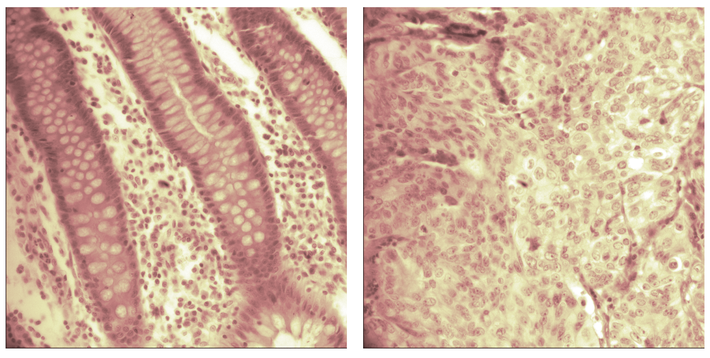
Overview

The classification of normal and malignant colon tissue cells is crucial to the diagnosis of colon cancer in humans. Given the right set of feature vectors, Support Vector Machines (SVMs) have been shown to perform reasonably well for the classification. In this paper, we address the following question: how does the choice of a kernel function and its parameters affect the SVM classification performance in such a system? We show that the Gaussian kernel function tuned with an optimal choice of parameters can produce high classification accuracy.

Motivation : Colorectal/bowel cancer is the 3rd most commonly diagnosed cancer in the UK after the lung and breast cancers, and remains the 2nd deadliest cancer disease after lung. In the UK alone, there were over 16,000 deaths due to bowel cancer in the year 2000. It is estimated that almost 80% of deaths caused due to colon cancer can be avoided if diagnosed at an early stage. However, the limited availability of pathological staff means difficulties in diagnosis due to high frequency of colon cancer incidents. In recent years, there has been growing interest in research of novel screening, diagnosis, and treatment methods.

Introduction : Hyperspectral imaging, which can capture 10s to 100s of spectral bands at varying wavelengths, have been traditionally used in remote sensing and related applications. The coupling of hyperspectral imaging with microscopy has found its way into biomedical applications, such as the classification of tissue cell in normal and malignant categories. Davis et al. proposed a supervised system for both segmentation and classification and achieved an accuracy of 86%. In our previously published work, unsupervised segmentation and supervised classification were employed achieving an accuracy of 87%. This paper presents the results of a study on improving the classification performance of a colon tissue cell classification using hyperspectral imagery.

Modality : Microscopic level image data cubes (using hyperspectral imaging sensors) of normal and malignant (adenocarcinoma) human colon tissue cells were acquired from archival H & E (Hematoxylin & Eosin) stained tissue sections. The dimensions of each data cube were 1024x1024x20, where 20 spectral bands in the wavelength interval 450-650nm were used. Figure 1 shows selected bands from two hyperspectral colon tissue cell image cubes containing benign and malignant cells.



(*a*) (*b*)

FIGURE 1. *Selected bands from hyperspectral colon tissue image data*

Contributions : We present the results of a study on improving the classification performance of a colon tissue cell classification system using hyperspectral imagery. Assuming that right set of feature vectors are employed by the system, we focus on finding optimal parameters for three kernel functions: linear, polynomial, and Gaussian.

Method

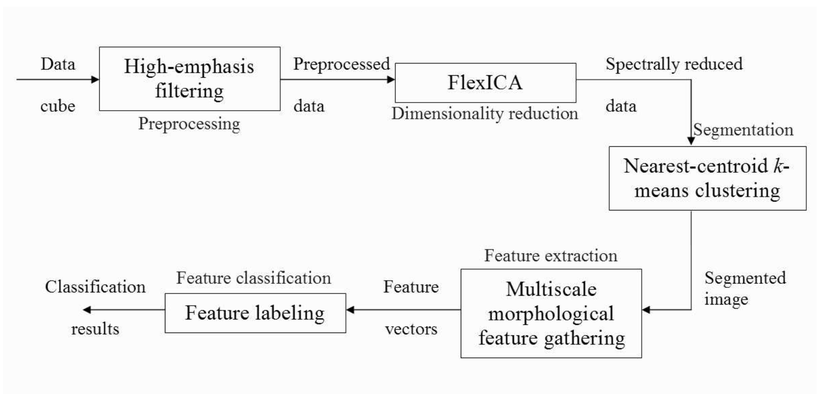


FIGURE 2. *Block diagarm of the classification algorithm*

1. *Pre-Processing* : A high-emphasis filtering of each spectral band of a data cube is carried out to force its data distribution towards heavy-tailedness.

2. *Dimensionality Reduction* : A variant of ICA sensitive to kurtosis is employed to separate out 4 'independent' components from the 20-dimensional spectral data.

3. *Segmentation* : Nearest centroid k-means clustering is performed on the extracted components (reduced data cube) to segregate it into cellular constituents: nuclei, cytoplasm, lumen, and lamina propria.

4. *Feature Extraction* : Multiscale morphological features (area, eccentricity, equivalent diameter, Euler number, extent, orientation, solidity, major axis length, minor axis length) are collected to exploit local as well as global discriminant characteristics.

5. *Feature Classification* : Multiscale morphological features are used for the training or testing of the SVM classifier.

SVM Optimization

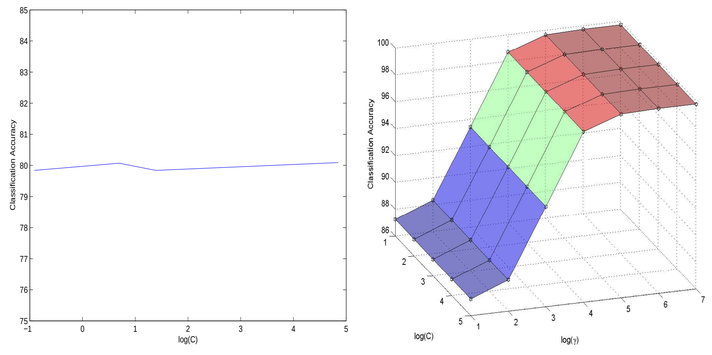
In SVM, the process of determining the decision boundary is greatly influenced by the selection of the kernel function, which in turn is heavily dependent upon the value of its free-parameters. In addition, each of the kernel functions has varying number of parameters (Table 1) that need to be tuned.

Kernel	Expression	Parameters
Linear	$< x_i, x_j >$	C
Gaussian	$e^{-\gamma\|x_i - x_j\|^2}$	γ, C
Polynomial	$(\gamma < x_i, x_j > + a)^d$	γ, C, a, d

TABLE 1. Typically used SVM kernel functions and their parameters

For a given application, it is hard to determine in advance which kernel function or set of respective parameters will produce the best results. Parameter selection is essentially the search for optimal parameters for a particular kernel. We employ a simple grid-search procedure which is iterative, but not always exhaustive depending upon the grid resolution.

Grid-search is the testing of various parameters values (taken from grid-points) for the SVM kernel and evaluating them to search for the best one. The feature data collection was divided into a training (1/4th) and testing (3/4th) portions. Figure 3 shows progressive results of parameter search of evaluating SVM on the testing data.



(*a*) Linear kernel

(*b*) Gaussian kernel: The effect of γ and C

(*c*) Polynomial kernel: C and a kept constant

(*d*) Polynomial kernel: γ and d kept constant

FIGURE 3. *Grid-search results for kernel parameters*

Experimental Results

The classification performance of the SVM using a Gaussian kernel approaches 99% at $\gamma = 17$ and $C = 1$. Quantitative results for this particular configuration are given below.

Accuracy (%)	Specificity (%)	Sensitivity (%)
99.72	99.62	99.82

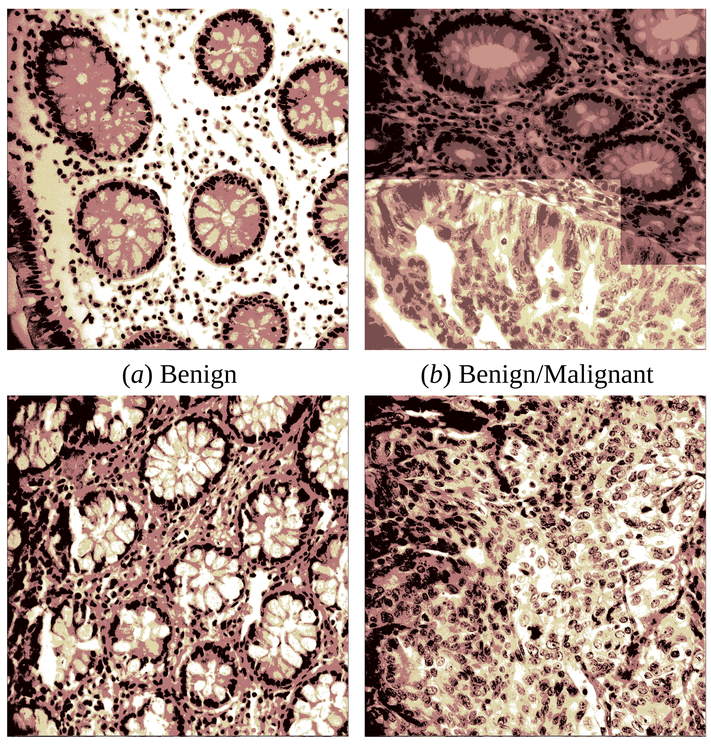
TABLE 2. Classification results with optimal parameters for Gaussian kernel

Discussion: These results show the promise exhibited by the SVM classifier and highlight the importance of selecting the right kernel and optimal set of kernel parameters. The fact that a Gaussian kernel outperforms linear and polynomial kernel settings may be due to a number of reasons:

a) It can determine a non-linear decision boundary.

b) It has fewer parameters than the polynomial and is consequently simpler to tune.

c) It faces less numerical difficulties (polynomial kernel value may go to infinity).



(*a*) Benign (*b*) Benign/Malignant

(*c*) Benign (*d*) Malignant

FIGURE 4. *Classification results*

Summary

Conclusions:

- In this paper, parameter selection procedure was studied to optimize the SVM classifier performance for a hyperspectral colon tissue cell classification system.
- It was shown that considerably high classification accuracy can be achieved for our tissue cell classification system by selecting optimal set of parameters for the Gaussian kernel.
- One of the limitations of our optimization approach is that it is rather exhaustive.

Future Directions:

- Our future work will look into efficient methods for the automatic selection of optimal parameters for the Gaussian kernel.
- The results of this initial study need to be validated on a larger dataset.

References

1. *Bowel Cancer Factsheet*, Cancer Research UK, April 2003
2. G. Davis, M. Maggioni, R. Coifman, D. Rimm, R. Levenson, 'Spectral/Spatial Analysis of Colon Carcinoma,' *United States and Canadian Academy of Pathology Meeting, Washington DC*, March 2003
3. K. Rajpoot, N. Rajpoot, 'Hyperspectral Colon Tissue Cell Classification,' *SPIE Medical Imaging*, February 2004
4. C. Hsu, C. Chang, C. Lin, 'A Practical Guide to Support Vector Machines,' *Technical Report, Department of Computer Science & Information Engineering, National Taiwan University*, July 2003