SVM Optimization for Hyperspectral Colon Tissue Cell Classification

Kashif Rajpoot[1] and Nasir Rajpoot[2]

[1] Faculty of Computer Science & Engineering, GIK Institute, Pakistan
[2] Department of Computer Science, University of Warwick, UK
kmr@giki.edu.pk, nasir@dcs.warwick.ac.uk

Abstract. The classification of normal and malginant 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 [4,13]. 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 combined with an optimal choice of parameters can produce high classification accuracy.

1 Introduction

Bowel cancer is the third most commonly diagnosed cancer in the UK after lung and breast cancer. It is the second most common cause of cancer death after lung cancer accounting for over ten percent of all cancer deaths. In the UK alone, there were over 35,000 colorectal cancer (a combined term for colon/rectum cancer) cases in the year 1999 and more than 16,000 deaths from bowel cancer in year 2000 [1]. The limited availability of specialist pathological staff and the huge amount of information provided by hyperspectral sensors means that user fatigue is a significant obstruction in the examination of these images and the identification of colon cancer in early stages. It is estimated that 80% of the deaths can be avoided if the cancer can be caught at its early stage. New improved screening and diagnosis methods could potentially save thousands of lives each year.

Hyperspectral imaging captures tens to hundreds of spectral bands at varying wavelengths in response to an image scene. The availability of this large amount of information can potentially help in the analysis of a scene. The use of hyperspectral images is widespread in remote sensing and related applications. The coupling of hyperspectral imaging with microscopy [10] has found its way into biomedical applications, such as the classification of colon tissue cells [4,13] into normal and malignant cells. Figure 1 shows selected bands from two hyperspectral colon tissue cell image cubes containing normal and malignant cells. In [4], Davis et al. proposed a completely supervised system for both segmentation and classification and achieved an accuracy of 86%. In our previous work [13], unsupervised segmentation and supervised classification were employed, with an

C. Barillot, D.R. Haynor, and P. Hellier (Eds.): MICCAI 2004, LNCS 3217, pp. 829–837, 2004.

increased potential of operating without significant human intervention, achieving an accuracy of 87%. Unfortunately, for both cases, the classification accuracy is not as high as desired in a real-world application of these algorithms.

In this paper, we present our work on improvement of the classification performance of the algorithm in [13]. Assuming that the right set of feature vectors were used to train and test the SVM classifier, we focus our attention in this paper on finding optimal parameters for three kernel functions: linear, Gaussian, and polynomial. A grid-search based method is employed in order to find an optimal set of parameters for each of the kernels. Our experiments show that the Gaussian kernel is most efficient in approximating the non-linear decision boundary between the two cell classes. Classification accuracy of over 99% was achieved using optimal parameters for the Gaussian kernel on a limited data set.

In the next section, a brief description of the classification algorithm of [13] is presented. Section 3 presents succint details of the SVM classifier optimization procedure. Experimental results are provided in Section 4, and the paper concludes with remarks on the effect of parameter selection on the classifier performance and some future directions.

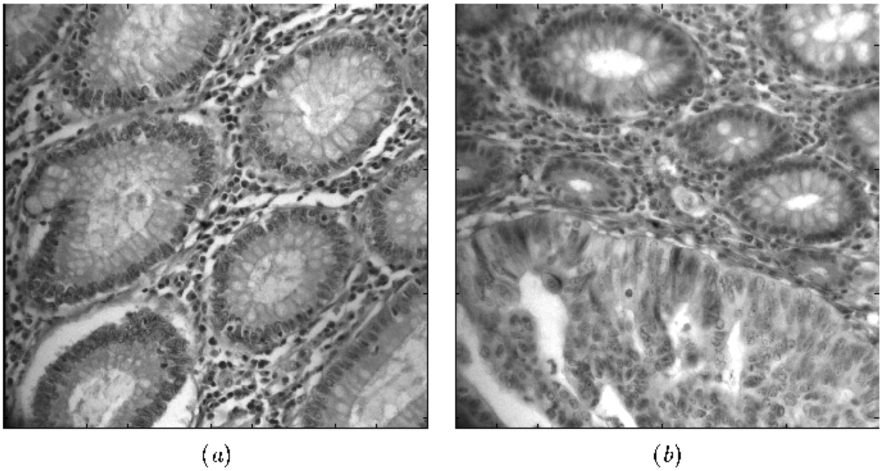


Fig. 1. Selected bands from hyperspectral colon tissue imagery. Two colon tissue sample images at 490nm; images contain (*a*) normal cells and (*b*) normal and malignant (towards the bottom-left) cells.

2 Materials and Methods

Microscopic level image data cubes of normal and malignant (adenocarcinoma) human colon tissue were acquired from archival H & E (hematoxylin & eosin) stained micro-array tissue sections. The dimensions of each data cube were 1024×1024×20, where 20 spectral bands in the wavelength interval 450–640nm were used. The challenge is the automated analysis of hyperspectral colon tissue images to classify between normal and malignant tissue sections with a reasonable accuracy. This will lead to a method that can be used without significant human intervention, once the machine is trained, and may be adopted as an assistance tool for the pathologists. Such a tool can be potentially helpful in evaluating the proportions of normal and malignant parts in an input colon image. A tissue cell classification problem can typically be approached through a traditional pattern recognition methodology. This involves: (i) segmentation, (ii) feature extraction, and (iii) classification. In hyperspectral imagery, a preprocessing step of dimensionality reduction may be included before the segmentation process. This approach is dissimilar to the one normally employed in the remote sensing field for classification problems, which merely exploits spectral signature for a direct classification task [11,6]. In this section, we give a very brief description of the classification method proposed previously by the authors. A more detailed treatment can be found in [13].

2.1 Segmentation

The segmentation of hyperspectral colon tissue images into four constituent parts of the human colon tissue cell (ie, nuclei, cytoplasm, lamina propria, and lumen) at the microscopic level was performed as follows. Independent Component Analysis (ICA) was employed to extract statistically independent components from the high-dimensional data. A preprocessing step of high-emphasis preceded the FlexICA variant [5] of ICA, which was used to achieve dimensionality reduction. The objective of this preprocessing was to force the data distribution towards heavy-tailedness, which is further exploited by the FlexICA algorithm, that is sensitive to kurtosis (4th order statistic). The extracted independent components (with reduced dimensionality in the spectral dimension) were fed into an unsupervised nearest-centroid (k-means) clustering algorithm, which resulted in a 1024×1024 labelled image for each hyperspectral image cube.

2.2 Feature Extraction

The segmented image was used to extract discriminant features which were subsequently utilized during the SVM classifier training stage. Multiscale morphological features (area, eccentricity, equivalent diameter, Euler number, extent, orientation, solidity, major axis length, minor axis length) were collected to extract the structural characteristics corresponding to each distinct 16×16 size patch of the segmented image. In almost all our experiments, morphological features performed better than statistical features mainly due to the fact that these were gathered from the segmented tissue cell image. The features associated with each patch were those for the patch itself and all of its parent resolutions up to a resolution of 256 × 256 (doubling the resolution each time) in order to exploit both local as well as global characteristics. This formed a 180-dimensional multiscale feature vector (36 features for each resolution) for every 16 × 16 patch of the image, yielding $4,096$ feature vectors for each hyperspectral image cube.

2.3 Classification

A total of $45,056$ such feature vectors were gathered through eleven hyperspectral image cubes. During the training stage of the SVM, about two-thirds $(30,000)$ of the feature vectors were used as training set while the rest $(15,056)$ were kept for a future testing stage. The algorithm achieved a classification accuracy of 87% on the unknown test data set using the Gaussian kernel. The need to improve the accuracy was one of the motivating factors for the investigation of SVM kernel optimization studied in this paper.

3 SVM Optimization

SVM [15] is an emerging area of research in the fields of machine learning and pattern recognition. SVM performs particularly well with high dimensional feature vectors and in case of lack of training data, two factors which may significantly limit the performance of most neural networks. Its true potential is highlighted when the classification of non-linearly separable data becomes possible with the use of a kernel function, which maps the input space into a possibly higher dimensional feature space in order to transform the non-linear decision boundary into a linear one. There exists a range of kernel functions, where a particular function may perform better for certain applications. The kernel functions can sometimes be categorized as *local* kernels (Gaussian, KMOD) and *global* kernels (linear, polynomial, sigmoidal) where local kernels attempt to measure the proximity of data samples and are based on a distance function rather than dot-product based global kernels. Table 1 lists the kernel expressions and corresponding parameters. Note that $< x, y >$ represents dotproduct, where x and y denote two arbitrary feature vectors. The process of determining the decision boundary is greatly influenced by the selection of kernel. In addition, each of the kernel functions have varying number of *free* parameters which can be selected by the *teacher*. As can be seen from Table 1, the performance of an SVM using linear, Gaussian, or polynomial kernels is dependent upon one, two, and four parameters respectively. All the kernels share one common parameter C, the constant of constraint violation which observes the occurring of a data sample on the wrong side of the decision boundary. Parameter γ of the Gaussian kernel denotes the width of the Gaussian radial basis function. For the polynomial kernel, d, γ, a respectively denote degree of the polynomial, coefficient of the polynomial function, and the coadditive constant. For a given application, it is hard to determine in advance which kernel function or set of respective kernel parameters will produce the best results. Selection of optimal parameters is currently a research issue in itself and is also known as the parameter or model selection problem. Another research direction is to make the parameter estimation process internal to the SVM classifier, with some researchers focusing on how to incorporate the process as an internal task to the SVM classifier, like finding optimal decision and margin boundaries.

Parameter (model) selection is essentially the search for optimal parameters for a particular kernel function. A simple way of doing so is the gride-search [7] procedure which is not always an exhaustive method (depending on the grid resolution). Other automatic methods, such as [14,2,9], exist but are iterative and can be computationally expensive too. Some other techniques which can be used to possibly improve the classification performance are: (i) *feature selection* [16]: to discard irrelevant features which may not be helpful for the classification process, (ii) *kernel selection*: choice of a kernel function for SVM for the mapping procedure, (iii) *data reduction* [12]: to discard irrelevant training samples as only samples near to the decision/margin boundary are important to the SVM; this is normally done by using a grid-search, (iv) *cross-validation*: splitting the data set into subsets to avoid overfitting and to improve on its general performance, (v) *marginal boundary determination*: finding an optimal boundary such that opposite class samples are well-separated, and (vi) *SVM classifier ensemble*: grouping or fusion of various SVM classifiers with different parameter settings.

4 Experimental Results and Discussion

Our experiments focussed on the selection of optimal parameters for each of the kernel functions. In order to avoid overfitting and to estimate the generalized performance, a 4-fold cross-validation exercise was conducted, where the whole feature data set was divided into four subsets such that one subset was iteratively tested using the classifier trained on the remaining data subsets. Table 2 shows classification accuracy results for all three kernels for different cross-validation trials. The parameter values used for these trials are: $C = 1, \gamma = 1, d = 3, a = 1$. As can be seen from the table, for a given kernel function, the classification accuracy does not vary significantly among different trials. This indicates that the optimal parameters are fairly generic in their application to classification of the unseen data. The search for optimal set of parameters can normally be carried out through an extensive experimentation process known as grid-search [7], which is the testing of different parameter values for the SVM kernels. This may

Table 1. Commonly used SVM kernel functions and their parameters

Kernel	Expression $K(x_i, x_j)$	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, d, a

Table 2. Classification accuracy (%) with 4-fold cross-validation

Trial#	Linear	Gaussian	Polynomial
1	82.2	86.9	83.1
2	81.5	87.8	83.2
3	81.5	86.7	83.9
4	82.0	87.1	83.7

Results of cross-validation with fixed parameters for all three kernels while the data was divided into four subsets and classification trials were carried out with the SVM trained on three subsets and tested on the fourth subset.

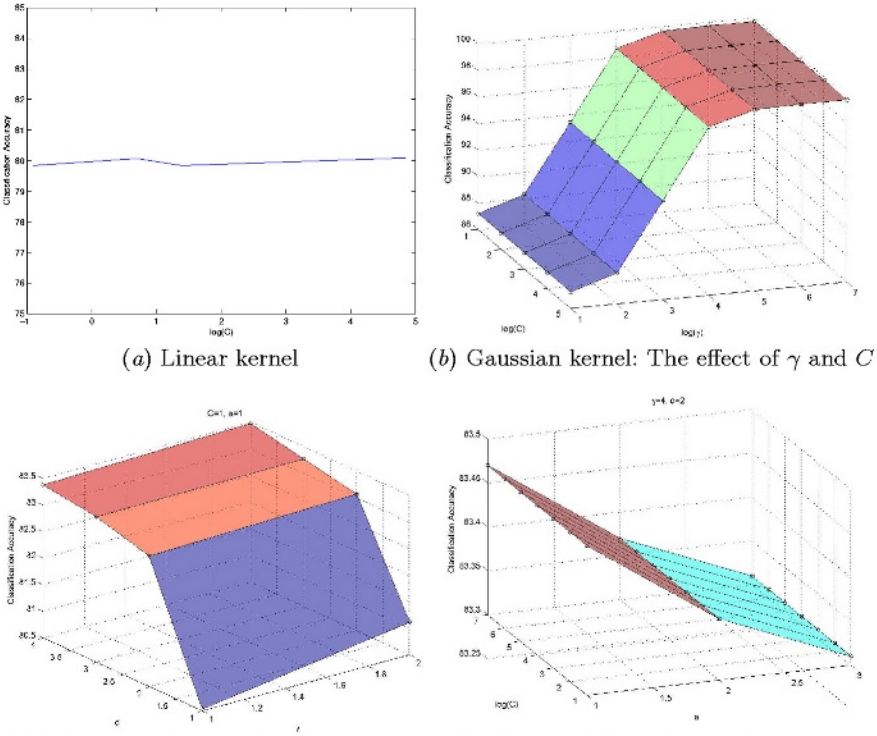
sometimes be preceded by a data reduction [12] preprocessing step to discard irrelevant data items, which are far away from the decision/margin boundary, in order to reduce the computational time involved in the search process. We omitted the data reduction stage since only one quarter of the total data samples were used for training and a coarse resolution grid-search based method was employed.

The feature data were divided into a training set ($11,000$ samples) and a test set ($34,056$ samples) and the search for optimal set of parameters was conducted using a grid-based method for all three kernel functions. Although Hsu et al. [7] suggest a grid range and grid steps of their choice for performing the grid-search, there is no hard and fast rule on this. Figure 2 shows progressive results of optimal parameter search for linear, Gaussian, and polynomial kernels. It can be seen from the Figure that a change in the value of parameter C, the penalty parameter common to all the kernels, does not have significant effect on the classification accuracy for any of the kernels, provided the remaining parameters are kept constant. It can also be observed from Figure 2(b) that the classification performance of the SVM using a Gaussian kernel approaches 99% for $\gamma = 17$ and $C = 1$. Results of classification are shown in Figure 3, where high contrast points to malignant sections and low contrast to normal sections. Quantitative results for this particular configuration of the SVM, ie using a Gaussian kernel function with optimal parameter values, are shown in Table 3. 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.

Table 3. Classification results (%) with optimal parameters for Gaussian kernel

Classification Accuracy	True +ve	True -ve	Sensitivity (Recall)	Specificity
99.72	99.62	99.82	99.82	99.62

According to [8], the hold-out testing (by distributing the available data into training and test sets) and quantified measures of Table 3 are a good way to estimate the generalization performance, though not totally unbiased, of the trained machine. The fact that a Gaussian kernel outperforms linear and polynomial kernel settings may be due to a number of reasons: (i) it can determine a non-linear decision boundary (not possible for a linear kernel), (ii) it has fewer parameters than the polynomial kernel and is consequently simpler to tune, and (iii) it faces less numerical difficulties (polynomial kernel value may go to infinity).



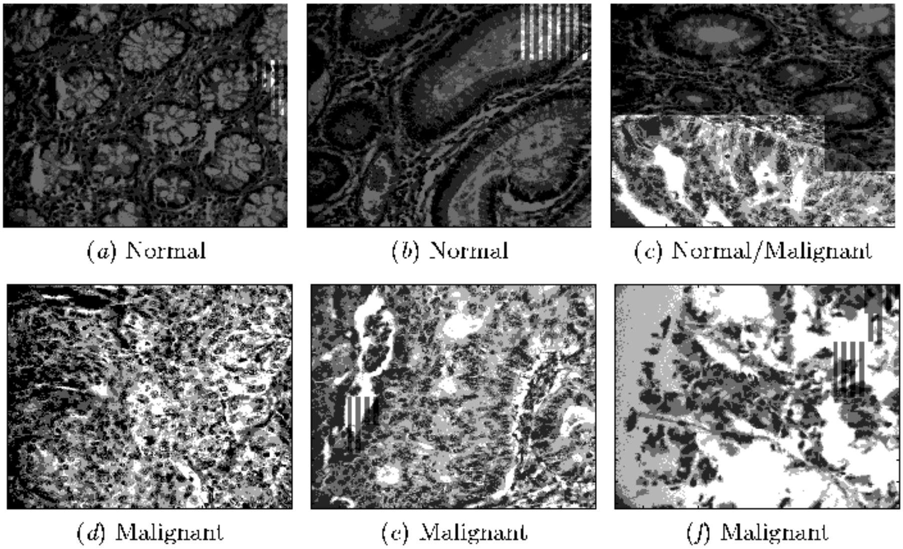
(*a*) Linear kernel (*b*) Gaussian kernel: The effect of γ and C

(*c*) Polynomial kernel: C and a constant (*d*) Polynomial kernel: γ and d constant

Fig. 2. Grid-search results for kernel parameters. The effect of varying the values of kernel parameters on the classification performance: (*a*) accuracy vs. parameter C for linear kernel, (*b*) surface of accuracy parameterized by C and γ for Gaussian kernel, surface of accuracy for polynomial kernel parameterized by (*c*) d and γ while keeping C and a constant, and (*d*) C and a while keeping d and γ constant. Where used, log is to the base 2.

5 Conclusions

In this paper, we studied parameter selection procedure to optimize the SVM classifier performance for a hyperspectral colon tissue cell classification system



(*a*) Normal (*b*) Normal (*c*) Normal/Malignant

(*d*) Malignant (*e*) Malignant (*f*) Malignant

Fig. 3. Experimental results for colon tissue cell images. Classification results for some colon tissue images overlaid on the original image (one of the spectral bands) showing 16 × 16 patches of cell areas classified as *normal* in *low* contrast and those classified as *malignant* in *high* contrast; on-off type of patchy artifacts can be observed in areas where the classifier perhaps does not have enough information.

[13]. 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. These results are in conformance with the recent findings [3] that the Gaussian kernel function is close to the natural diffusion kernel which produces the best mapping results. Gaussian kernel is also more efficient compared to the other two kernels, since it is a local distance-based kernel. One of the limitations of our optimization approach is that it is rather exhaustive. Our future work will look into efficient methods for the automatic selection of optimal kernel parameters for the Gaussian kernel and validation of our results on a larger data set.

Acknowledgements. The authors gratefully acknowledge obtaining hyperspectral imagery data used in this work from and having fruitful discussions with Ronald Coifman and Mauro Maggioni of the Applied Mathematics Department of Yale University. We also thank Abhir Bhalerao, Roland Wilson, and anonymous reviewers for their useful suggestions to improve the quality of this paper.

References

1. Bowel cancer factsheet, April 2003. Cancer Research UK.
2. O. Chapelle and V. Vapnik. Model selection for support vector machines. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, 1999.
3. R. Coifman and S. Lafon. Geometric harmonics. Technical report, Department of Applied Mathematics, Yale University, 2003.
4. G. Davis, M. Maggioni, R. Coifman, D. Rimm, and R. Levenson. Spectral/spatial analysis of colon carcinoma. In *United States and Canadian Academy of Pathology Meeting, Washington DC*, March 2003.
5. R. Everson and S. Roberts. Independent component analysis: A flexible non-linearity and decorrelating manifold approach. In *Proc. IEEE NNSP*, 1998.
6. J.A. Gualtieri and R.F. Cromp. Support vector machines for hyperspectral remote sensing classification. In *Proc. SPIE Workshop, Advances in Computer Assisted Recognition (ACAR)*, October 1998.
7. C.W. Hsu, C.C. Chang, and C.J. Lin. A practical guide to support vector machines. Technical report, Department of Computer Science & Information Engineering, National Taiwan University, July 2003.
8. J. Joachims. Estimating the generalization performance of a svm efficiently. In *Proc. Intl. Conf. on Machine Learning*. Morgan Kaufmann, 2000.
9. J.H. Lee and C.J. Lin. Automatic model selection for SVM. Technical report, Department of Computer Science & Information Engineering, National Taiwan University, November 2000.
10. R. Levenson and C. Hoyt. Spectral imaging & microscopy. *American Laboratory*, pages 26–33, November 2000.
11. G. Mercier and M. Lennon. Support vector machines for hyperspectral image classification with spectral based kernels. In *Proc. IEEE Intl. Geoscience & Remote Sensing Symposium (IGARSS)*, 2003.
12. Y.Y. Ou, C.Y. Chen, S.C. Hwang, and Y. J. Oyang. Expediting model selection for SVM based on data reduction. In *Proc. Intl. Conf. on Systems, Man, and Cybernetics (ICSMC)*, October 2003.
13. K.M. Rajpoot and N.M. Rajpoot. Hyperspectral colon tissue cell classification. In *SPIE Medical Imaging (MI)*, February 2004.
14. C. Staelin. Parameter selection for support vector machines. Technical report, HP Labs, Israel, 2002.
15. V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.
16. J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for SVMs. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2000.