Effectiveness of Spectral Band Selection/Extraction Techniques for Spectral Data

Marina Skurichina, Sergey Verzakov, Pavel Paclik, and Robert P.W. Duin

Information and Communication Theory Group, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, P.O. Box 5031, 2600GA Delft, The Netherlands
m.skurichina@tudelft.nl

Abstract. In the past few years a variety of successful algorithms to select/extract discriminative spectral bands was introduced. By exploiting the connectivity of neighbouring spectral bins, these techniques may be more beneficial than the standard feature selection/extraction methods applied for spectral classification. The goal of this paper is to study the effect of the training sample size on the performance of different strategies to select/extract informative spectral regions. We also consider the success of these methods compared to Principal Component Analysis (PCA) for different numbers of extracted components/groups of spectral bands.

1 Introduction

Densely sampled spectral measurements became a standard tool in many applications such as medical diagnostics or industrial quality control. The amount of data to be dealt with has increased even further due to widespread adoption of hyperspectral imaging sensors capturing spectral readings in spatial raster. The acquired spectral information is, however, largely redundant due to low intrinsic dimensionality of the studied phenomena. Therefore, raw spectral measurements are usually reduced for the sake of data transmission, visualization or data analysis. In this paper, we discuss a specific type of data reduction techniques targeting supervised pattern classification.

The examples of successful methods to find discriminative spectral regions are an Optimal Region Selector (ORS) [1] guided by a genetic algorithm, a top-down and bottom-up multiresolution feature extraction algorithms proposed by Kumar et al. [2], Recursive Band Selection (RBE) [3] etc. The advantage of these techniques is that they make use of the connectivity between neighbouring spectral bins when finding discriminative groups of spectral bands, while the standard feature reduction approaches (such as forward/backward feature selection or PCA [4]) neglect the apriori available information on the ordering of spectral wavelengths. Spectral band selection techniques are also preferred to standard feature reduction techniques, because they allow us to find discriminative regions in spectra instead of single bands or "generalized" features (like in PCA). By this, specialists can make the relation between the informative group of spectral bands found and the physical background of a studied phenomenon. It also implies the possibility to design cheap devices to perform

D.-Y. Yeung et al. (Eds.): SSPR&SPR 2006, LNCS 4109, pp. 541 – 550, 2006.

measurements only for few spectral regions that make sense for discrimination instead of measuring spectra for a wide range of all possible emission wavelengths.

Sometimes it is not possible to find clear discriminative spectral regions especially for spectral data representing mixtures of materials. The information useful for discrimination might be spread over all (or over the majority) spectral features. However, in the case of highly dimensional data with a relatively small amount of available measurements, it is needed to reduce the data dimensionality in order to construct a reliable classification rule [5]. Then PCA may be used, as it insures that all information contained in original features is preserved in extracted principal components. But the first few components (describing the largest variance in the data) do not guarantee the best discrimination between data classes, because PCA is an unsupervised feature extraction technique that does not make use of data class information. For a better classification performance, one may need a larger number of principal components.

Both, the spectral band selection techniques and PCA, have benefits and drawbacks that may depend on the training sample size, the number of desirable components/regions, and on the type of the spectral data they are applied to. What concerns the spectral band selection methods, their success depends on many factors: the exact strategy to find spectral regions, the criterion to select best regions, a merging function to produce a singe value introducing the group of spectral bands and finally the classification rule used to evaluate the success of feature extraction. We can expect that for PCA and the spectral band extraction methods, small training sample sizes may cause problems to find good discriminative components/regions. The PCA may be imperfect when a too small number of principal components is considered. The spectral band extraction methods may tend to select single bands (representing noise in the data) when they are forced to find a large number of spectral regions.

The goal of this paper is to compare the classification performances of different spectral band selection strategies (that extend standard feature selection techniques by using the spectral ordering information) mutually and to PCA for different training sample sizes and different numbers of extracted components/regions. Two real data sets representing two-class problems are used in our study. They are described in section 2. Different feature selection/extraction strategies used to find discriminative spectral regions are introduced in section 3. The results of our experimental study are discussed in section 4. Conclusions are summarized in section 5.

2 Data

Our study is performed on two real-world datasets representing two-class problems.

The first dataset consists of the autofluorescence spectra acquired from healthy and diseased mucosa in the oral cavity. The measurements were performed at the Department of Oral and Maxillofacial Surgery of the University Hospital of Groningen [6]. The spectra were collected from 97 volunteers with no clinically observable lesions of the oral mucosa and 137 patients having lesions in oral cavity. The measurements were taken at 11 anatomical locations with excitation wavelength 365 nm. After preprocessing [6] each spectrum consists of 199 bins. In total, 581 spectra representing healthy tissue and 123 spectra representing diseased tissue were obtained

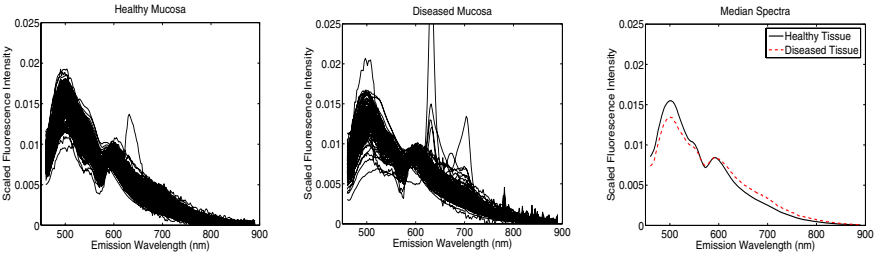

Fig. 1. Normalized autofluorescence spectra for healthy and diseased mucosa in oral cavity

after a thorough inspection of the database and removing all doubtful measurements. In order to reduce a large deviation in a spectral intensity within each data class, spectra were normalized by the unit area. Normalized autofluorescence spectra of healthy and diseased tissues and their median spectra are illustrated in Fig. 1.

The second dataset represents histograms of DNA content of tumour cells in a breast tissue [7]. The data are provided by the Pathology Department of De Wever Hospital of Heerlen. The DNA of all cells in a breast tissue sample is stained with a fluorchrome that emits red light after irradiation with a laser beam. The emitted light photons are collected in a photo multiplier tube in the flowcytometer and converted to electrical pulses that are proportional to the amount of DNA in the cells. After counting 20000 cells, a histogram is made of the DNA content of these cells. Each histogram is described by 256 wavelength channels of flowcytometer. After removing the first two and the last two histogram bins (which contain only noise), each histogram consists of 252 bins. The dataset contains 448 histograms of DNA content representing aneuploid breast tumour cells and 199 histograms describing DNA content of diploid breast tumour cells. The histograms are normalized by the unit area. The examples of these histograms and the median histograms are presented in Fig. 2.

For our experiments, training data sets with 10, 50 and 100 objects per class are considered for both datasets studied. Each time the training objects are chosen randomly from the total set. The remaining data are used for testing. The prior class probabilities are set to be equal as the data are very unbalanced and the real prior class probabilities are unknown. To evaluate the performance of diseased tissue diagnostics when different feature selection/extraction methods are used, we have

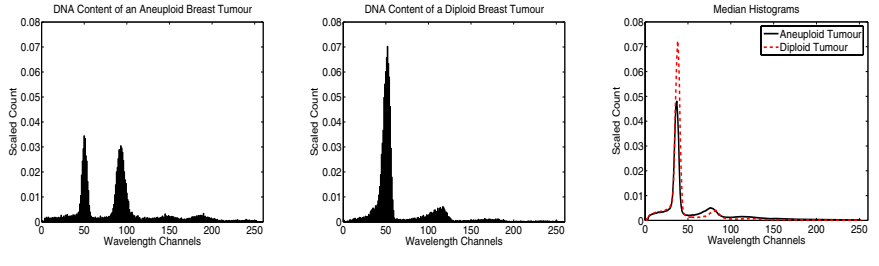

Fig. 2. The median histograms and selected examples of normalized histograms of the DNA content of 20000 cells obtained for aneuploid and diploid breast tumours

chosen the Regularized Linear Classifier (RLC) [8] which constructs a linear discriminant function assuming normal class distributions and using a joint class covariance matrix for both data classes. The value of the regularization parameter used is equal to 10^{-8}. All experiments are repeated 20 times on independent training sets. In all figures the averaged results over 20 trials are presented. The standard deviation of the reported mean generalization errors (the mean per two data classes) is about 0.01 for each considered case.

3 Strategies for Spectral Band Selection

As a rule, each of spectral band selection techniques proposed in the literature uses another criterion to find the discriminative spectral regions and a different function to merge a group of spectral bands into a single value representation. The choice of such a criterion or a merging function can seriously affect the performance of a spectral band selection technique. In order to eliminate the influence of these two factors on the classification performance of the studied spectral band selection strategies, we use the same criterion and the same merging function for all of them.

As a discriminant measure (criterion) to evaluate a discriminative capacity of extracted spectral regions, we use the Mahalanobis Distance (MD) between data classes:

$$MD = (\mu_A - \mu_B)'(p\Sigma_A + (1-p)\Sigma_B)^{-1}(\mu_A - \mu_B), \quad (1)$$

where μ_A, μ_B and Σ_A, Σ_B are the means and the covariance matrices of data classes A and B, respectively; p is the prior probability of the data class A. The larger Mahalanobis distance, the larger discriminative capacity between data classes.

A merging function used by us to reduce the dimensionality of each considered spectral region to a single value representation is the mean function which simply takes the average of spectral intensities in the region.

In our study we consider the following spectral band extraction strategies.

Approach 1. *GLDB-TD.* A top-down multiresolution feature extraction algorithm proposed by Kumar et al. [2], partitions the original p-dimensional spectra into smaller subspaces by using a top-down recursive algorithm. First, the best place to split spectra into two parts is found by computing a discriminant measure between data classes. The criterion value obtained on the parent space is compared with the criterion values calculated on the children subspaces. If the child subspace has a higher discrimination than the parent space, then it is partitioned further. We stop the partitioning, when no child subspace shows an improvement in its discrimination capacity compared to the parent space. The GLDB-TD algorithm is fast, but the final set of spectral regions found is suboptimal, because the optimization is performed only in one-dimensional way: a discrimination capacity is evaluated for each spectral region separately but not for a total set of selected spectral regions.

Approach 2. *GLDB-BU.* A bottom-up generalized local discriminant bases algorithm proposed by Kumar et al. [2], merges p original bands in larger subspaces by using a bottom-up recursive algorithm. First, the best pair to merge among all possible pairs of neighbouring single bands is found by computing a discriminant measure between

data classes. The criterion value obtained on the best merged band is compared with the criterion values calculated on its component subspaces. If the merged space has a higher discrimination than the component subspaces, the merge is accepted and we move to the next level. Elsewise, the merge is denied and we consider the second best pair to merge on this level. If no merge is found that gives the better discrimination than component subspaces, then the merging procedure stops. This strategy has the same limitation as GLDB-TD: the optimization is performed only for one spectral group.

Approach 3. *Recursive Band Elimination (RBE).* The RBE technique proposed by Verzakov at al. [3] is a modification of the SVM shaving technique. We apply RBE using the Regularized Linear Classifier (RLC) instead of SVM in order to compare this strategy with other spectral band selection techniques in even conditions. First, the linear classifier is trained on the original p-dimensional spectral data. The absolute values of RLC coefficients $|w_i|$, $i = \overline{1,p}$, are used to split spectra into an initial set of spectral regions. Namely, minima of $|w_i|$ are the splitting points to obtain groups of spectral bands. Then each of S obtained spectral regions is merged into a single feature and the Recursive Feature Elimination (RFE) [9] is applied to perform a backward elimination of spectral regions. At each step of RFE, we train the RLC on the features representing groups of spectral bands and remove the spectral region corresponding to the smallest absolute value of the RLC coefficients.

Both, RBE and GLDB-BU, recursively reduce the number of spectral regions. GLDB-BU starts from p single band regions and merges them iteratively (no spectral band is omitted) until the discrimination cannot be improved anymore by merging the spectral regions. The RBE starts from $S < p$ spectral regions defined by the coefficients of RLC and eliminates them one by one till one last spectral region remains. The RBE performs multivariate optimization for the spectral region selection. The absolute values of RLC coefficients (instead of Mahalanobis Distance) are used as a criterion to select discriminative groups of spectral bands.

Approach 4. *Sequential Partitioning (SP)* [10]. It also performs multidimensional optimization for the spectral region selection. First, spectra are partitioned into two parts by finding the best split (with the optimal criterion value over all possible partitions) in the space of two features extracted from the two spectral regions. When the first split location is anchored, we look for the second optimal split in such a way that the criterion value in a three-dimensional space (on three features extracted from the three spectral regions) is the largest over all possible locations for the second split. We fix the second split location and repeat the procedure until the desired number of spectral regions S is found. In this approach, all spectral bands are used in the partitioning of spectra. However, some of them may be uninformative - introducing only noise. It is good to remove them, as they may deteriorate the classification when they are included in the extracted spectral regions. One way to do this is described below.

Approach 5. *Sequential Partitioning and Elimination of uninformative spectral bands (SPE)* [10]. After a desired number of spectral regions S is found by the previous approach, we shrink the spectral regions removing uninformative bands. We proceed in a sequential way (region by region) moving from the most left region to

the most right one. In order to shrink the spectral region, we consider all possible subregions of the reduced size in the region and find the subregion with the largest criterion value in S-dimensional space (where one feature represents a shrunk subregion of the spectral region under consideration and the rest S-1 features are extracted from the other S-1 spectral regions which definitions are fixed for a moment). After shrinking the first spectral region, we anchor its new definition and move to the next spectral region in order to exclude uninformative spectral bands. This method does not guarantee the optimal shrinking for all regions in general, because it is highly dependent on the proceeding order of spectral regions.

Approach 6. *Sequential Selection (SS)* [10]. The discriminative spectral regions are selected sequentially one by one. At each step s ($s = \overline{1, S}$) we consider all possible region definitions (of arbitrary size) in spectra. For each definition we calculate the discriminant measure in s dimensions: one feature represents a current potential pretender for the most discriminative spectral region and other s-1 features are extracted from the previously selected spectral regions. The region (a potential pretender) with the largest criterion value in s-dimensional space is picked as the most discriminative spectral region (in combination with the s-1 previously found optimal regions). In this approach the overlapping and non-overlapping spectral regions may be selected. Some spectral bands might be not selected at all to participate in spectral regions.

Approach 7. *Sequential Selection of Non-overlapping discriminative spectral regions (SSN)* [10]. This approach is identical to SS but the overlapping spectral regions are not allowed to be selected.

Approach 8. *Floating Partition (FP)* [11]. First, spectra are uniformly partitioned to a predefined number of spectral regions S. At each step, we allow the borders between spectral regions to float one spectral bin aside from the current position. Among 3^S possible mutations we select the partition that provides the highest discrimination according to the selected criterion. We repeat the procedure until no improvement in discrimination capacity can be found. This method performs multivariate optimization by simultaneously adjusting all spectral regions. However, it is still a suboptimal procedure because the drifting step for region borders is limited to one spectral bin. The efficacy of this method can be improved by enlarging the drifting step d. But this leads to computational burdens because one has to rank $(2d+1)^S$ cases at each step of the procedure. We could apply this approach upto 10 spectral regions (with d=1) at most.

4 Experiments and Discussion

Before studying the benefits of extracting/selecting the discriminative spectral regions in the comparison with the PCA approach for different training sample sizes N and for different numbers of extracted components/regions B, let us make few remarks on the datasets used. When measuring autofluorescence spectra of healthy and diseased tissues in oral cavity, in reality the autofluorescence spectra of the mixture of materials (skin, tissue under skin, bone and saliva) are obtained. The useful information for lesion diagnostics is hidden in overlapping peculiarities of different materials. The

clear-cut discriminative spectral regions do not exist. The useful information for lesion diagnostics is spread over the whole spectrum. What concerns the histograms of the DNA content, the main information is concentrated around the two largest peaks in the left part of the histogram and in the region between them (see Fig. 2). The peaks and the region in between describe different phases of the cell division cycle. The amount of the DNA in these phases characterizes different types of tumour cells. So, all useful discriminative information is concentrated in spectral bands around these peaks of the histogram. Thus, our two datasets represent two extreme cases: when no separate discriminative regions exist (in autofluorescence spectra) and when we have a few well-defined discriminative regions (in histograms). For the first dataset, we can expect that the PCA may be very effective because the principal components aggregate the information represented in all spectral bands. For the second dataset, the spectral band selection techniques, that select regions around the histogram peaks, will be the most beneficial.

When considering results obtained for the autofluorescence spectra (see left plots in Fig.3), we see that for small training sample sizes (N=10+10), all techniques perform similarly with a slight preference for RBE, SSN and FS (the last one only for two extracted spectral regions). The training data are not enough representative to find correctly the discriminative spectral regions. Due to a limited number of training objects, only 19 principal components can be extracted by PCA. When increasing the training sample size, all techniques perform better, but the relative advantages of their generalization errors do not change much. The exceptions are the SP and SPE strategies, which become the best among the studied spectral band selection techniques for large numbers of extracted spectral regions when the training sample size is large (N=100+100). In agreement with our prior knowledge on the dataset, the more regions/components are selected, the better all methods perform. The most successful technique is PCA with a large number of principal components that accumulate useful information spread over the whole spectrum. The PCA is followed by the SP strategy which uses all spectral bands in spectral partitions. Excluding some spectral bands (in SPE) deteriorates the classification performance. However, in order to get a medical insight of the studied phenomenon, for data compression (in remote sensing) or for building filters in the sensor, we are interested in finding few discriminative spectral regions rather than many of them. The PCA is unsupervised feature extraction technique that finds directions in a feature space with the largest variance that are not necessarily discriminative. The spectral band selection methods are supervised techniques that take advantage of data class information. By this, they outperform PCA for small numbers of extracted components/regions. Interestingly, the RBE strategy was the best when 8-15 spectral regions are selected. Usually RBE converges to its best solution for a relatively large number of spectral regions. However, for this dataset, the clear preference for particular discriminative regions does not exist. Probably, RBE outperforms all other strategies due to the superior discriminant measure used (the absolute values of RLC weights are used instead of MD).

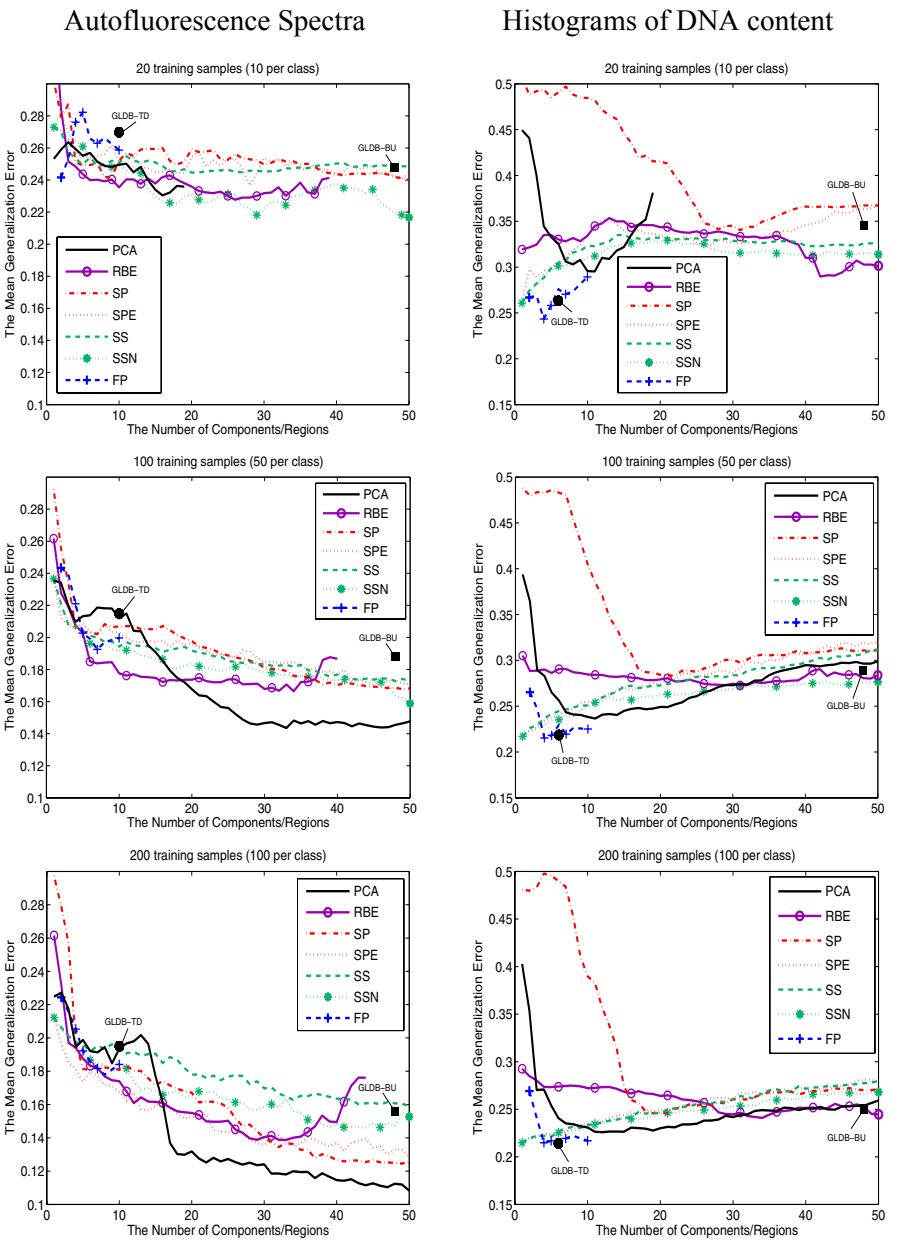

Fig. 3. The mean generalization error (GE) of LDA for training sample sizes 10, 50, and 100 objects per class, when different methods are used to select discriminative spectral regions for autofluorescence spectra measured in oral cavity (*left plots*) and for histograms of DNA content in tumour cells (*right plots*). Because the GLDB-TD and GLDB-BU algorithms terminate automatically using a data-driven criterion, only a single point is given in each plot. The standard deviation of the mean GE is around 0.01.

For histogram data (see right plots in Fig. 3), the performance of all strategies is improved by increasing the training sample size. But the general mutual behaviour of the generalization errors for all methods remains the same. The performance of all techniques worsen when the number of extracted regions/components grows after 8-10 regions. It is logical because mainly the spectral regions around and between the peaks (which are related to the cell division cycle) provide useful information for the classification of tumour cells. Adding more regions is equivalent to adding noise and cannot improve the classification. Indeed, the best results are found by the spectral band selection techniques when less than five spectral regions are retrieved. The spectral band selection strategies SPE, SS, SSN and GLDB-TD perform equally nice and the best among all studied techniques. The first three techniques select the spectral regions around the DNA content peaks, finding the most discriminative parts of spectra. GLDB-TD algorithm can also make a successful split of spectra converging at five-six spectral regions on average. However, the sequential partitioning (SP) completely fails for small numbers of extracted spectral regions. It happens by two reasons. First, all bands (also uninformative) are kept in spectral regions (when uninformative spectral bands are eliminated in SPE, the performance is drastically improved). Second, the partitioning in SP is done in a sequential way. Once the split is found, it cannot be adjusted anymore. The best split found for partitioning into two regions might be very far from the optimal one for partitioning into more regions than two. The FP strategy overcomes this problem by simultaneously adjusting all spectral regions. It competes with other spectral band selection techniques when spectra are split into four-five spectral regions. As the first five principal components extracted by PCA are not discriminative, its performance is poor. Since both, RBE and GLDB-BU, are bottom-up recursive procedures for finding informative spectral regions, they usually converge to the suboptimal solution at a relatively large number of spectral regions (around 45 for RBE and around 95 for GLDB-BU). Hence for this problem (with three-four discriminative spectral regions by definition) the performance of RBE and GLDB-BU is worse than the performance of other feature selection techniques (with exception of SP).

5 Conclusions

The success of spectral band extraction techniques varies over the potential of the spectral data depending on how information useful for classification is introduced: locally (in a few clear-cut discriminative spectral regions) or globally (spread over the majority of spectral wavelengths). The supervised spectral band selection techniques which make use of the connectivity of spectral wavelengths in spectral data (one-dimensional ordering) are more beneficial than unsupervised PCA when one needs to find a small number of discriminative spectral regions/components. However, which spectral band selection technique is preferred seems to be defined by the problem and the criterion used to select the best regions. These issues need more study in the future.

References

1. Nikulin, A., Dolenko, B., Bezabeh, T., Somorjai, R.: Near Optimal Region Selection for Feature Space Reduction: Novel Preprocessing Methods for Classifying MR Spectra. NMR in Biomedicine **11** (1998) 209-216
2. Kumar, S., Ghosh, J., Crawford, M.M.: Best-Bases Feature Extraction Algorithms for Classification of Hyperspectral Data. *IEEE* Transactions on Geoscience and Remote Sensing **39** (2001) 1368 - 1379
3. Verzakov, S., Paclik, P., Duin, R.P.W.: Feature Shaving for Spectroscopic Data. Lecture Notes in Computer Science, Springer-Verlag, Vol. 3138 Berlin Heidelberg New York (2004) 1026-1033
4. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press (1990)
5. Jain, A.K., Chandrasekaran, B.: Dimensionality and Sample Size Considerations in Pattern Recognition Practice. In: Krishnaiah, P.R., Kanal, L.N. (eds.): Handbook of Statistics, Vol. 2. North-Holland, Amsterdam (1987) 835-855
6. De Veld, D.C.G., Skurichina, M., Witjes, M.J.H., et.al.: Autofluorescence Characteristics of Healthy Oral Mucosa at Different Anatomical Sites. Lasers in Surgery and Medicine **32** (2003) 367-376
7. Verzakov, S., Duin, R.P.W.: The Tangent Kernel SVM for Calibration-Stable Histogram Discrimination. Proceedings of the ASCI conference, ASCI, Delft (2005) 73-80
8. Friedman, J.H.: Regularized Discriminant Analysis. JASA **84** (1989) 165-175
9. Guyon, I., Weston, S., Barnhill, S., Vapnik, V.: Gene Selection for Cancer Classification Using Support Vector Machines. Machine Learning **46**(13) (2002) 389-422
10. Skurichina, M., Paclik P., Duin, R.P.W, et.al.: Selection/Extraction of Spectral Regions for Autofluorescence Spectra Measured in the Oral Cavity. Lecture Notes in Computer Science, Vol. 3138 Springer-Verlag, Berlin Heidelberg New York (2004) 1096-110
11. Meloni, S.: Finding Discriminative Bands in Auto-Fluorescence Spectra for Automatic Cancer Diagnosis. Master Thesis, Cagliari University, Sardinia, Italy (2004)