Feature Shaving for Spectroscopic Data

Serguei Verzakov, Pavel Paclík, and Robert P.W. Duin

Information and Communication Theory Group
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology
Mekelweg 4, 2628 CD Delft, The Netherlands
{s.verzakov,p.paclik,r.p.w.duin}@ewi.tudelft.nl

Abstract. High-resolution spectroscopy is a powerful industrial tool. The number of features (wavelengths) in these data sets varies from several hundreds up to a thousand. Relevant feature selection/extraction algorithms are necessary to handle data of such a large dimensionality. One of the possible solutions is the SVM shaving technique. It was developed for applications in microarray data, which also have a huge number of features. The fact that the neighboring features (wavelengths) are highly correlated allows one to propose the SVM band-shaving algorithm, which takes into account the prior knowledge on the wavelengths order. The SVM band-shaving has a lower computational demands than the standard SVM shaving and selects features organized into bands. This is preferable due to possible noise reduction and a more clear physical interpretation.

1 Introduction

In pattern recognition, objects are usually described by a number of features. Not all of them are equally informative for the problem at hand. Therefore, feature selection is an important step in solving a classification problem. It simplifies the classification task offering a faster and cheaper solution and, moreover, it allows to improve the classification performance by avoiding the curse of dimensionality. Feature selection methods can be divided into two groups: a) univariate approaches, where each single feature is tested for its ability to discriminate between classes and b) multivariate approaches, where all features are ranked according to some criterion, which takes all of them into account at once. Univariate approaches are simple to implement, but multivariate approaches give better results as they take into account the feature dependences.

Recently, a number of feature selection methods under the name of 'shaving' have been developed. Among them there are the SVM shaving [1] and the PCA shaving [2, 3]. Shaving approaches are similar but not equivalent to the backward feature elimination technique. In general, shaving algorithms remove a small portion of features at each step based on some criterion calculated on all the features available at that moment. During the backward feature elimination, an importance of each feature is estimated according to a criterion calculated on the

A. Fred et al. (Eds.): SSPR&SPR 2004, LNCS 3138, pp. 1026–1033, 2004.

feature set present at that moment but this particular feature. The backward feature elimination algorithm should theoretically provide better results than shaving. Yet, the shaving approach is much faster, while it is still able to provide a good solution.

Originally, shaving methods were applied to the microarray data. It has been shown that shaving techniques are useful in finding a small group of features (genes) significant for discrimination in high dimensional microarray data [1–3]. In our research, we are dealing with the problem of spectra classification. Based on the number of channels in the spectrometer, the feature sizes may vary from several hundreds up to a thousand. It is natural to try to apply these methods to the high-resolution spectral data. However, there is an important difference between the spectral data and gene arrays. In spectroscopy, neighboring features (wavelengths) are often highly correlated (more than 90 percent). It makes shaving methods first detect these correlations and only then the actual feature selection starts. In the case of small training set, the estimation of the local correlations may be very imprecise. This leads to a waste of the computational time and it may also result in loss of important features. The outcomes of shaving methods are the sets of original features. However, it is more natural for spectroscopy to select continuous bands of wavelengths and derive (possibly weighted) average representative feature from each band. Such features are easily interpretable from the physical point of view and also more robust to a change of the measurement device.

Another family of feature extraction/selection approaches under the name of GLDB was proposed in [4]. There, the neighboring wavelengths are combined into one feature based on log-odds class posterior probabilities (top-down approach) or on a product of the Fisher criterion and correlation between the features (bottom-up approach). Although these methods take features dependencies into account, all the criteria are applied to the each single region of wavelengths. Thus this family of methods is not fully multivariate.

In this paper, we propose a use of modification of the SVM shaving algorithm, *SVM band-shaving*, which makes use of specific properties of spectral data. Briefly, we combine neighbouring wavelengths into bands at first and then apply the shaving algorithm to them. The paper is organized as follows. In the section 2 we shortly describe the SVM shaving algorithm and our modification of it. Then, in section 3 we present the results of numerical experiments and we summarize with a short conclusion in section 4.

2 Shaving Algorithms

All shaving algorithms rely on a computation of the ranking vector $w \in R^d$, where d is the number of features. The absolute value of the element $w(i)$ estimates the importance of i-th feature e.g. for a discrimination task. After removing the least important feature or some portion of such features, the algorithm recalculates the weight vector for the reduced feature set. Recalculation on the data set with reduced dimensionality is desirable or even necessary, since the

algorithm starts from the complete set of features, for which the estimated importance values can be very imprecise due to the curse of dimensionality. The algorithm continues the reduction of feature set size until the specified number of features is reached. It is also possible to estimate the classification error at each step on the current feature set and choose the one which offers a good tradeoff between the small number of features and an acceptable classification error.

It is possible to use as a ranking vector the weight vector w of a linear classifier

$$f(x) = sgn((w, x) + b) \tag{1}$$

In [1], the usage of the SVM classifier was proposed due to its reputation of being robust to the curse of dimensionality and being able to provide better estimations at early steps of shaving.

In [1], the SVM shaving was applied to microarray data. The relations between gene expression levels are either unknown or very complicated. In our case, we have extra prior information about spectra: the order of the features (wavelengths) is meaningful. The spectral values of the neighboring features are typically highly correlated (of course, this is only true for data with a sufficiently high spectral resolution). We use the word *connectivity* to name this property of spectral data sets.

The use of any additional information about a data set is similar to (but more specific than) regularization and in a similar way can help to reach better generalization abilities. In this article, we propose a modification of the SVM shaving technique which takes into account connectivity in the feature set. First we combine features into continuous groups (bands). For this purpose, we use absolute values $|w(i)|$ of the weights $w(i)$ obtained by training linear SVM on all the features. The bands are separated from each other by local minima of $|w(i)|$. To find local minima we estimate the first and the second derivates of $|w(i)|$ using Savitsky-Golay filter [6] with the second order polynomials. By averaging data in each band we create a new feature set to which the standard SVM shaving algorithm will be applied. The small number of features in the new feature set allows us to remove them one by one instead of removing some percentage of them. In all cases we use the ν-SVM algorithm [5], because its parameter ν has a more convenient interpretation (the estimation of the classification error) than the parameter C of the standard C-SVM algorithm.

As input parameters of the algorithm, additionally to the ν parameter of SVM, the minimum number of bands (stopping criterion) and the size of smoothing window for Savitsky-Golay filter should be specified. The selection of the meaningful minimum number of bands is a matter of the experiment and can be only roughly estimated beforehand e.g. as the number of significant principal components. It is worth to notice that classes can overlap substantially for the small numbers of bands which leads to the long execution times of SVM routines. All these problems are also present in the standard SVM shaving. The size of smoothing window should be selected as a largest interval on which $|w(i)|$ (or spectra themselves) can be well fitted by the second order polynomials for any i.

The pseudo-code of the proposed algorithm is presented below:

Input:
the training data set D,
the complete feature set F,
the parameter ν of the ν-SVM classifier,
minimum number of bands min_bn,
maximum smoothing window size max_ws.

Output:
the sequence of feature sets $F_1 \supset ... \supset F_n$.

Algorithm:

1. Calculate the weight vector w for the full feature set F using ν-SVM algorithm.
2. Calculate the absolute values of the weight vector elements $w_a(i) = |w(i)|$.
3. Find the set of bands $B = \{b_1, ..., b_m\}$ which are separated by the minima in $w_a(i)$. Use Savitsky-Golay [6] algorithm with the second order polynomials and with the smoothing window less or equal to max_ws to estimate the first and the second derivatives of w_a.
4. Create a new feature set F_1 such that each feature $z(i)$ is a signed mean value of features $x(j)$ which belong to the band b_i.

$$z(i) = \frac{1}{|b_i|} \sum_{j \in b_i} sgn(w(j)) * x(j) \tag{2}$$

5. Perform the standard SVM shaving (using ν-SVM algorithm) on F_1 removing each time one band producing the sequence of feature sets $F_1 \supset F_2 ...$ until no more than min_bn bands left.

One can use a validation data set to estimate the classification error on the resulting sequence $F_1 \supset ... \supset F_n$ to judge which subset has the smallest number of bands while yielding a suitable performance.

3 Numerical Experiment

For a demonstration of our algorithm we use the data from the CD of [7]. This is a 191-channel airborne multispectral scanner data set which contains a hyperspectral image of Washington DC Mall. The sensor system used in this case measured a response in 0.4 to 2.4 μm region of the visible and infrared spectrum. The task is to discriminate between seven classes of pixels: Roofs, Roads, Paths, Trees, Grass, Water and Shadows. We demonstrate our algorithm on the Roofs/Paths classification. For our calculations we have selected 30 spots of each

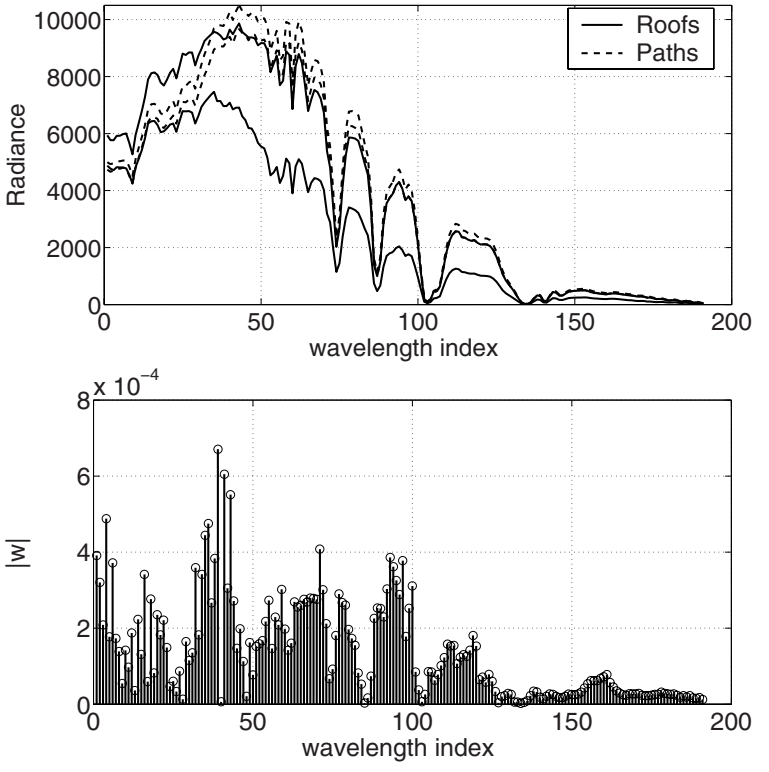

Fig. 1. Upper plot: lower and upper quartiles (25% and 75% levels of a distribution) for each wavelength for both classes. Radiance units are arbitrary. Bottom plot: absolute values of the weights $w(i)$ resulting from training of the SVM classifier on a feature set containing all 191 wavelengths.

class. Each spot consists of 9 pixels. The spots were manually selected to guarantee a representative examples and placed faraway from each other. We used a 5-fold cross-validation to estimate classification errors. The parameter $\nu = 0.05$ of ν-SVM algorithm was selected after a few probe runs and proved to be a good choice. The values of ν greater than $\nu = 0.05$ lead to larger classification errors due to the insufficient penalizing of the classification error. At smaller values, the solutions of the SVM problem start to show early signs of overtraining with the decreasing of the number of features in the shaving procedure. This happens, because in low dimensional data sets classes start to overlap, so unreasonably high penalization of margin errors (ν is much smaller than Bayes error) leads to narrow margin and a bad generalization ability. See [8] for more details.

In Fig. 1, the spectra of both classes are shown, as well as the result of SVM applied to the non-reduced feature set. After a few experiments, we have selected the upper limit for the smoothing window $max\_ws = 11$. The result of the band extraction is shown in Fig. 2.

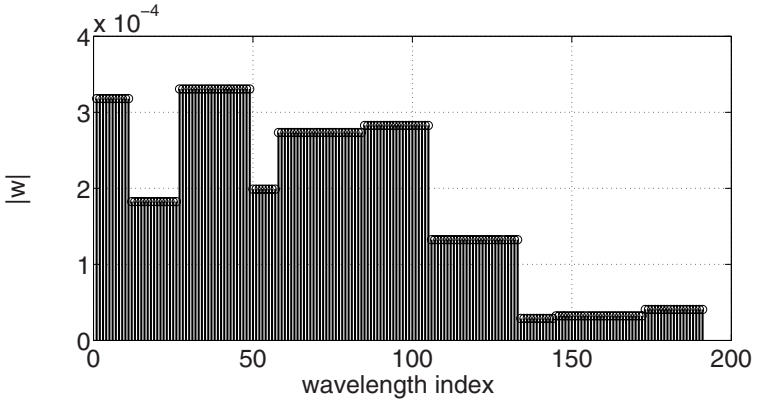

Fig. 2. Absolute values of the weights $w(i)$ after the step 4 of the SVM band-shaving algorithm (band extraction). The cumulative weight of each band is equally distributed among features from this band.

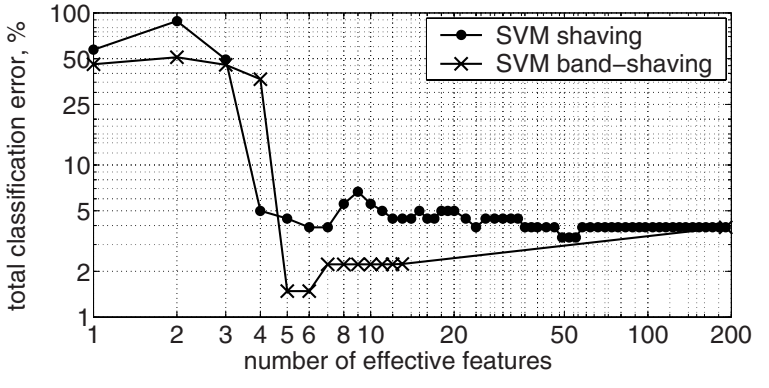

Fig. 3. The classification errors for the feature sets selected by shaving algorithms. The x-axis represents the number of effective features, i.e. the dimensionalities of spaces in which classifiers were trained and tested.

The total classification errors on the feature sets selected by the standard SVM shaving and the SVM band-shaving are shown in Fig 3. Both methods start from the same entire feature set (191 original features). Then, the standard procedure gradually removes the least important features by portions of 5% of the remaining features. The classification error remains almost the same during the shaving. It reaches minimum at 49 features. The number of features equal to 6 seems to be an optimal choice because of a significant dimensionality reduction (from 191 to 6 features) and still a low classification error.

The absolute values of weights $w(i)$ of the SVM trained on the selected 6 features are shown in Fig. 4. The classification performance for the number of features less than 4 is very bad due to overlap.

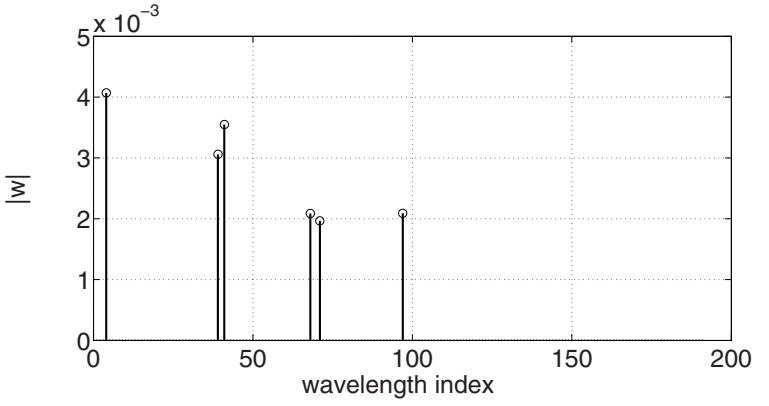

Fig. 4. Absolute values of SVM weights $w(i)$ which are the result of the training SVM classifier on a feature set containing 6 features selected by the SVM shaving procedure.

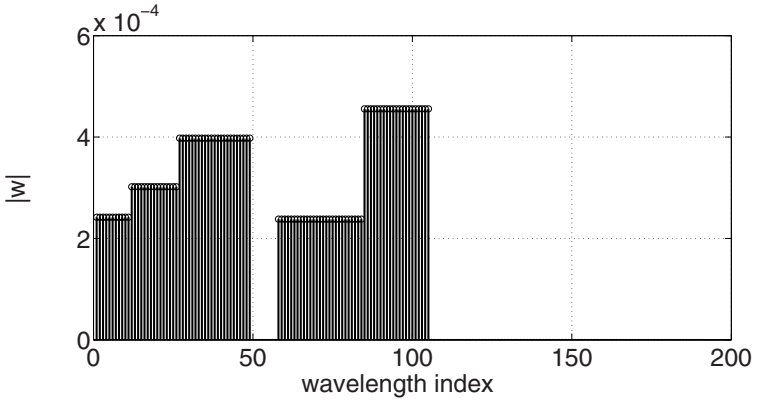

Fig. 5. The absolute values of SVM weights $w(i)$ trained on 5 bands selected by SVM band-shaving. The weight of each band is equally distributed among features from this band.

The SVM band-shaving immediately jumps from the entire feature set to a feature set containing only approximately 10 features. This number varies slightly in each cross-validation run. The results suggest that only by combining the features into the bands, the classification performance can be improved. Moreover, this performance can be better than one of a classifier trained on the same number of features selected by the standard procedure. During the removing the least important bands, the classification error becomes smaller. It reaches the minimum at the number of bands equal to 5 (see Fig. 5). At lower numbers of bands results show the clear signs of a substantial overlap between classes.

4 Conclusion

We proposed a variant of the SVM shaving algorithm, the SVM band-shaving, which takes into account the connectivity property of the spectra. The conducted experiment shows that our algorithm may outperform the standard SVM shaving technique. The SVM band-shaving removes the whole band at once. Thus the number of retrainings of the classifier is smaller than in the standard SVM shaving procedure. On the other hand, our algorithm requires the specification of an additional parameter: the maximum size of smoothing window. A few runs of the whole procedure or an expert knowledge on the nature of data are necessary to select a proper value of this parameter. It is also worth to mention that 5 bands selected by the SVM band-shaving contain in total about 90 original features (wavelengths). So the application of some band shrinking algorithm would be useful. Our results, although preliminary, are very encouraging. We plan to study these techniques further on and apply them to other data sets.

Acknowledgments

For our computations, we used OSU SVM by J. Ma, Y. Zhao, and S. Ahalt [9]. We would like to thank E. Pekalska for the helpful comments on the manuscript. This work was supported by the Dutch Foundation for Applied Sciences (STW).

References

1. I. Guyon, J. Weston, S. Barnhill and V. Vapnik: Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning **46**(13), 389-422, 2002.
2. T. Hastie, R. Tibshirani, et al.: Gene Shaving: a New Class of Clustering Methods for Expression Arrays. Technical report, Department of Statistics, Stanford University, 2000. `http://www-stat.stanford.edu/~hastie/Papers/`.
3. T. Hastie, R. Tibshirani, et al.: "Gene shaving" as a method for identifying distinct sets of genes with similar expression patterns. Genome Biology **1**(2), research0003.1–0003.21, 2000.
4. S. Kumar, J. Ghosh, J. and M. M. Crawford: Best-Bases Feature Extraction Algorithms for Classification of Hyperspectral Data. IEEE Transactions on Geoscience and Remote Sensing **39**(7), 1368-1379, 2001.
5. B. Schölkopf, A. Smola, R.C. Williamson, and P.L. Bartlett: New support vector algorithms. Neural Computation **12**, 1207-1245, 2000.
6. A. Savitzky and M. J. E. Golay: Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry **36**(8), 1627-1639, 1964.
7. D. Landgrebe: Signal theory methods in multispectral remote sensing. John Wiley & Sons, 2003.
8. V. Vapnik: The Nature of Statistical Learning Theory. Springer: New York, USA, 2000.
9. J. Ma, Y. Zhao, S. Ahalt: OSU SVM Classifier Matlab Toolbox (version **3.00**), Ohio State University, Columbus, USA, 2002.
`http://www.ece.osu.edu/~maj/osu_svm/`.