SPIE Vol. 2315, Image and Signal Processing for Remote Sensing, pp. 472-482, 1994

Filtering remote sensing data in the spatial and feature domains

Freddy Fierens and Paul L. Rosin

Institute for Remote Sensing Applications
Joint Research Centre, I-21020 Ispra (VA), Italy
email: {freddy.fierens,paul.rosin}@jrc.it

ABSTRACT

We present a comparative study of the effects of applying pre-processing and post-processing to remote sensing data both in the spatial image domain and the feature domain. We use a neural network for classification since it is not biased by *a priori* assumptions about the distributions of the spectral values of the classes. Spatial smoothing was applied both as pre- and post-processing steps. Pre-processing involved smoothing the image spectral values by means of anisotropic diffusion, whereas iterative majority filtering was applied as a post-processing step to improve spatial coherence by reclassifying pixels. While it is common practice to filter the image before classification (smoothing) or after classification (iterative majority filtering) it is less obvious what happens if pre-processing is applied to the training or image data in feature space. To minimise the effect of noisy training pixels we applied a k-nearest neighbour filtering algorithm to the training data. This involved reclassifying each training pixel by the majority class of the set of k closest training pixels (in terms of Euclidean distance) in feature space. The procedure eliminates isolated training pixels and tends to produce more compact class clusters. The effects of all spatial and spectral filtering methods were validated by applying them to three different testcases.

1 INTRODUCTION

The two main approaches for training multi-spectral remote sensing data classifiers are supervised and unsupervised learning. Supervised learning involves training the classifier with a learning set of pixels whose landuse label is known. Typically, the training set is assembled either by an on-site survey or by a human photo-interpreter using aerial and/or satellite photographs and maps. Assessment of the performance of the classifier can be done by classifying a separate test data set of pixels with known landuse labels. It is generally considered good practice to take the testing and training pixels from different areas in the image. Unsupervised learning on the other hand does not require labelled training pixels. Instead, this method tries to identify natural clusters in the data. The method however, does require the setting of parameters such as expected number of clusters, maximum cluster variance, etc. It is the task of the data analyst to afterwards identify the natural clusters in terms of landcover classes or class combinations. The main advantages of the unsupervised clustering scheme is that it does not force multimodal clusters into a unimodal distribution (such as can be the case for maximum likelihood if no precautions are taken), and that it is a more economic procedure since it does not require a training set. Supervised classification on the other hand has the advantage of classifying the images with labels that are meaningful for the end-users. As an additional advantage, it enables an assessment of the accuracy of the classification. However, accumulating the training set, e.g. by an on-site survey, can be very expensive.

Since the most common purpose of remote sensing image classification is the generation of a landuse map using predefined landuse classes, we will only discuss supervised classification. A problem with supervised classification is its reliance on ground truth training data, thus making it sensitive to any errors in this data set. Potential sources of error can be categorised as data acquisition errors and data preparation errors. Examples of possible data acquisition errors are the time lapse between the ground truth survey and the acquisition of the images. This results in errors in the spectral values for the training data because the ground truth may be out of date (e.g. a time lapse of one or two weeks may be sufficient to invalidate the landcover label in an agricultural area due to harvesting or ploughing). Different interpretations by different ground truth observers are another source of acquisition errors. Since usually entire fields or parts of for instance a wood are identified as single parcels with the same ground truth class, these areas will invariably contain mixed pixels that will confuse the learning. Once the ground truth data has been acquired, it has to be converted into a usable format for generating the final training and testing sets. This involves the alignment of ground truth polygons with corresponding areas in the image, extracting the spectral values from these image areas and combining them with the correct ground truth label. Depending on the classification system, further format conversion stages may be necessary. Inaccuracies in outlining the training areas on the image can cause boundary pixels with either incorrect labels or mixed class signatures to be introduced into the training set. Every step in the data preparation is a potential source for introducing errors. The spectral filtering technique we describe is one way of reducing the effects of these errors.

2 NEURAL NETWORKS AS FEATURE CLASSIFIERS

The supervised classifiers used in these experiments are multi-layer perceptron neural networks that take real-valued inputs. Each input node of the neural network takes the values of one spectral band, each output node represents one class. During the learning phase, the spectral values of the training pixels are presented to the input nodes of the network while the output node representing the class to be recognised is clamped to the active state. All other output nodes are clamped to the inactive state. The backpropagation algorithm then attempts to minimise the error between the desired classification and the actual classification performed by the network by adapting the weights between the network nodes. Once the global error is minimised for all the training pixels, the network is ready to classify other input data.

In our experiments, we choose neural networks rather than standard statistical classifiers because they make less *a priori* assumptions about class distributions.[2] More in particular, they do not assume classes to be unimodal or to be normally distributed. However, since neural networks are trained to minimise a least square error, they are very sensitive to outliers in the training data.[13] Although these outliers could be valid training points, the network might artificially attribute too much importance to them. This could lead to overfitting of the training data.

3 STANDARD DATA PREPARATION

Standard preparation of the training data for neural network learning involves the canonisation of the input data per band followed by a randomisation of the input pixels. Canonisation consists of rescaling the input training data so that the mean input value is zero and the standard deviation is one in all bands. Randomisation of the order in which the input training pixels are presented to the neural network is important in order to encourage faster convergence.

4 EXPERIMENTS

We performed five different experiments, using a combination of spatial and spectral filtering schemes applied to training, input and classified data.

4.1 Stratification of the training data

First we compared the effect of stratifying the training data with the use of non-stratified data. In the literature of neural networks applied to remote sensing, there does not appear to be any consensus on whether to use stratified or non-stratified data. In the experiments of Hepner[6] and Benediktsson,[1] stratified training data is used, whereas other experiments, notably those of Kanellopoulos[7] use non-stratified data. Stratified data ensures that a substantial number of training pixels are present, even for infrequently occurring classes, whereas the non-stratified data introduces a form of *a priori* probability for class occurrence. The latter assumes that the relative combined area of the training polygons for each class roughly represents the percentage of occurrence of that particular class in the scene. Note however, that it is also possible to explicitly incorporate *a priori* probabilities into a neural network during training.[12]

4.2 Spatial smoothing

Spatial smoothing of the image data is a standard technique, whether used to pre-process the raw image data or to post-process the assigned classification labels.[4] Since classification of parcels is performed on a per pixel basis then, assuming spatial coherence in the scene (i.e. it contains homogeneous regions), smoothing decreases the effect of unwanted local variations and thus decreases the probability of mislabelling. Simple spatial smoothing such as filtering the image through an averaging mask has the disadvantage of blurring edges. Therefore there has been much interest in edge-preserving smoothing which eliminates noise while retaining significant (edge) features. Examples of these techniques can be found in the k-neighbourhood averaging scheme[4] and the multiple mask filters of Nagao and Matsuyama.[8] Anisotropic diffusion is a relatively new and promising edge preserving smoothing technique that makes use of an optimisation approach.[11] The degree of smoothing in anisotropic diffusion is controlled by specifying a minimum allowable intensity contrast between neighbouring pixels. The diffusion effectively removes all lower contrast changes while preserving higher contrast changes. A small value of the smoothing parameter will only remove low-contrast detail. Larger values of the smoothing parameter will remove higher-contrast detail. In the other approaches the amount of smoothing depends more on the largest homogeneous intensity patch in the pixel neighbourhood. Here, the effect of increasing the window size is less clear since the amount of smoothing depends on a combination of the intensity contrast and the spatial size of the detail.

A different form of spatial smoothing can be applied as a post-processing step after classification. Some methods, such as iterative majority filtering,[14] do not rely on pixel intensities, but only on the spatial neighbourhood of the classified pixels. Alternatively, relaxation can be applied to the output probabilities of the pixel labels, rather than on the spatial context of class labels only.[4] Since these forms of spatial smoothing have the effect of increasing spatial coherence, not only the visual appearance but also the accuracy of the resulting classification may be improved by it. For smoothing applied as a post-processing step, we experimented only with iterative majority filtering.

4.3 Spectral smoothing

In the next set of experiments, we investigated the effect of filtering the training data in feature space in order to increase the spectral coherence in the training classes. This is analogous to increasing spatial coherence in an image by applying a smoothing filter. Spatial coherence assumes that pixels in a local neighbourhood in the image have similar attributes. These attributes can be spectral values or class labels. Likewise, spectral coherence assumes that the spectral characteristics of pixels belonging to the same class are similar. Both spatial and spectral coherence assume that outliers are undesirable. Even though outliers are not necessarily wrong, they can still have a negative effect on the overall classification accuracy. Overtraining can force neural netowrks to construct local boundaries around outliers, rather than generalising by absorbing them in the surrounding class. Increasing the spectral coherece of the feature space values decreases the likelihood that this problem occurrs.

Since data in feature space is not uniformly distributed and the values are nominal rather than ordinal, the data cannot be smoothed with a simple convolution kernel as is mostly used for spatial smoothing. The effect of smoothing in feature space can be obtained however, by applying a variation of k-nearest neighbour filtering. K-nearest neighbour filtering compares each pixel in the training set to its k nearest neighbours. If the class assigned to the central pixel is different from the majority of its k neighbours, this pixel can be considered an outlier. We can then either reassign this pixel to the majority class of its neighbours, or simply delete it from the training set. We investigated the effect of both approaches. This process of modifying the training data is related to the editing rules used by Hart,[5] Penrod,[10] Wilson[15] and Devijver.[9] Their motive however, was to improve the efficiency of the k-nearest neighbour classifier by removing pixels that would have little or no effect on the final result. Contrary to our approach, these editing rules have the effect of preserving outliers rather than removing them.

4.4 Combined spatial-spectral smoothing

Whatever techniques are used to improve the spectral separation in feature space, class overlap in the data to be classified imposes an upper limit to the accuracy that can be achieved. Further increases in classification accuracy are only possible by including the spatial context of the individual pixels into the procedure. The simplest method of achieving this is probably the use of texture masks which calculate extra features based on a local neighbourhood of the pixels. These extra features can in some cases differentiate between otherwise overlapping spectral classes.

We have examined two different approaches of integrating spatial contextual information in the classification process (see figure 1). Both methods start by k-filtering the training data in feature space. The smoothed training data is then used to train a neural network. The first method then smooths the input image intensities (for each band independently) using the anisotropic diffusion method mentioned above and then classifies the smoothed image. The second method first classifies the input image and then iteratively applies a cycle consisting of iterative majority filtering followed by k-nearest neighbour filtering of the classified pixels in feature space. The mapping from the iterative majority filtered image into feature space is a many to one mapping, mapping many points in the classified image that have the same spectral values but different class labels into the same position in feature space. In order to avoid loosing any information in the many to one mapping, the coordinates and labels of all the individual image points are retained. This ensures that the mapping of the k-nearest neighbour filtered pixels from feature space back into the classified image is a one to one mapping.

4.5 The effects of neural network complexity

The effects of the spatial as well as the feature space filtering operations cannot be completely separated from the effects introduced by neural network complexity. The number of nodes per hidden layer in a neural

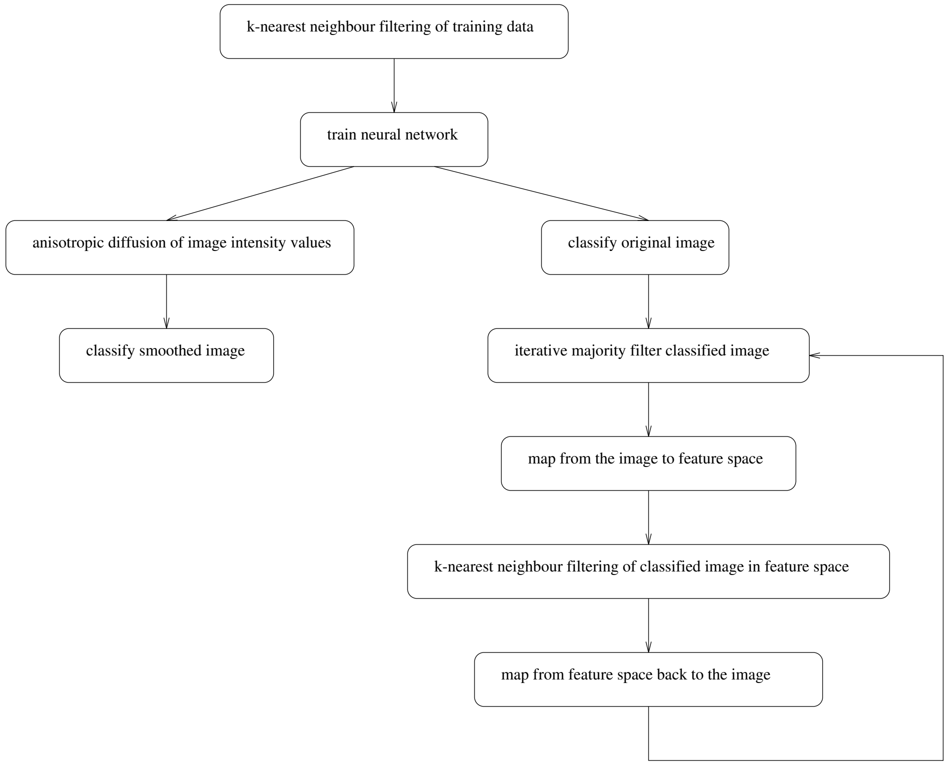


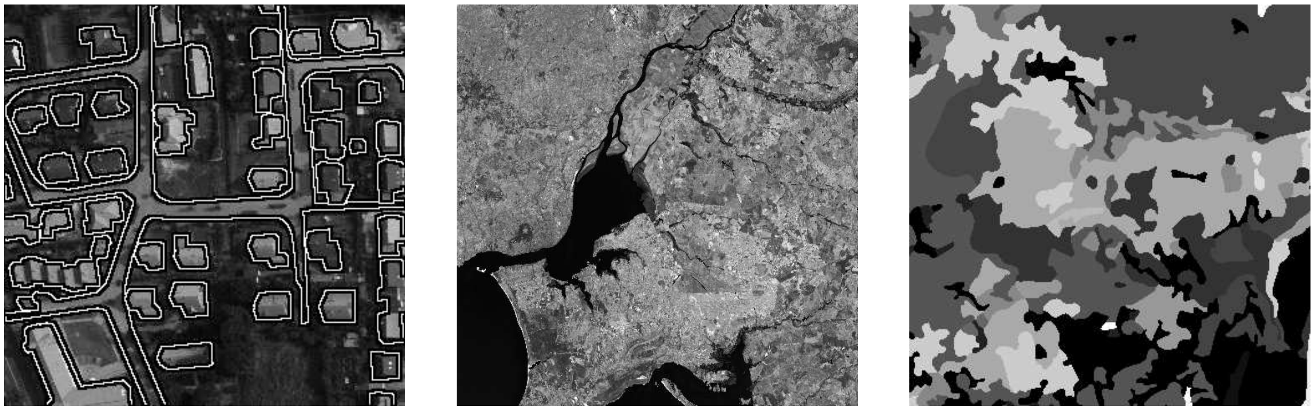
Figure 1: Integration of spatial and spectral filtering

network has an important influence on the complexity of the decision boundaries that can be formed. It can be expected that simplifying the shape of the training data clusters can have a positive effect for networks that are not capable of forming complex boundaries, but might otherwise have a negative effect on the classification accuracy of networks that do have this capability. On the other hand, outliers in the training data may cause a complex network to overfit, thus reducing the resulting total accuracy. K-nearest neighbour filtering in feature space, which has the effect of deleting or reassigning outliers, may reduce the probability of overfitting. In order to test the influence of these different parameters, we also performed a test in which neural network complexity was gradually increased in terms of the number of its hidden nodes, and investigated the classification accuracies for various values of the k-parameter in k-nearest neighbour filtering.

5 TESTCASES

As a first practical example to test the above mentioned principles, we used a 3-band RGB aerial photograph of a suburban neighbourhood in Glandorf, Germany. This photograph, together with ground truth data covering the same region was made available by the ISPRS Working Group III/3.

A second testcase was situated near the town of Lisbon in Portugal and consisted of a quarter 7-band Landsat-TM scene together with our own ground truth data gathered by an on-site survey. Since band 6 of a multi-spectral Landsat-TM image is a thermal band and has a different resolution compared to the other bands, we only used bands 1-5 and band 7. Of the original 22 classes identified by the ground truth survey, classes that gave rise to too much confusion were collapsed into superclasses. This resulted in a data set with 16 different class labels.



(a) Glandorf aerial image (b) Portugal Landsat-TM scene (c) CORINE map as ground truth

Figure 2: Images used in the testcases

A third testcase made use of a part of the same Landsat-TM image used in testcase 2. This time however, we used the CORINE landcover map as ground truth. In the area we considered, 22 different landcover classes were present (according to the CORINE map). Since the size of the area covered by some of the classes was too small, only 19 classes were retained.

Complete ground truth was available for the Glandorf and CORINE testcases. For the Portugal testcase, the ground truth data was obtained by the on-site survey and therefore restricted to small training polygons. Consequently, the effects of spatial filtering could not be tested on this data set.

6 RESULTS

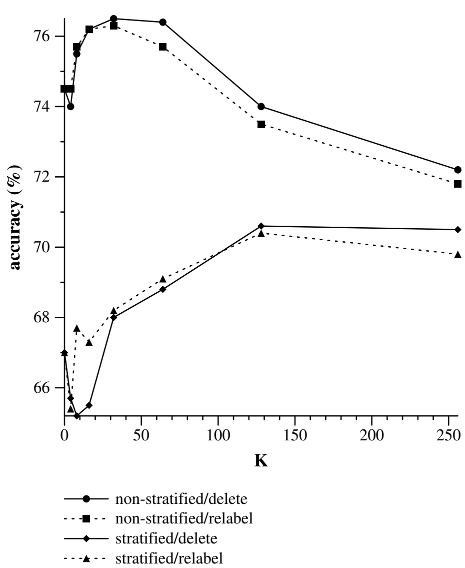


Figure 3: Glandorf testcase

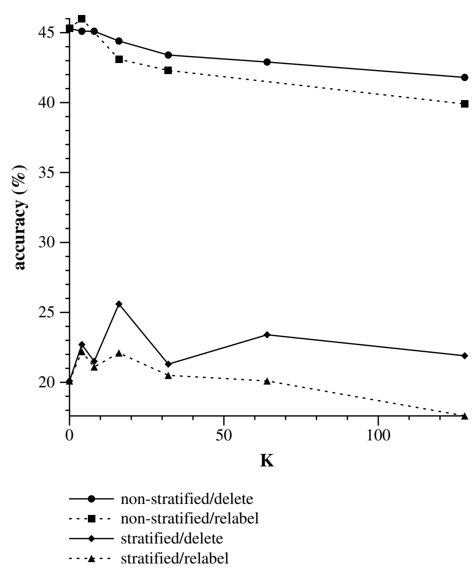


Figure 4: CORINE testcase

The results of the different experiments were plotted on graphs with the abscissa showing the different sizes of k-neighbourhood used in feature space filtering and the ordinate giving the resulting accuracy of the classified image.

Accuracy was measured in terms of percentage correct classification. Although more sophisticated accuracy measures exist (notably the KHAT statistic and the normalised accuracy measure of Congalton[3]), our simpler accuracy measure was sufficient to study the effect of the different experiments as we are only interested in general relative increase or decrease of accuracy.

The first experiment concerned stratification of the training data and showed that the use of non-stratified training data is the better approach. This conclusion was confirmed by both the Glandorf and CORINE testcases as is clear from figures 3 and 4. Stratification was not tested on the Portugal data set. From the results obtained we can conclude that it is advantageous if the training data set reflects the *a priori* probabilities or ratios of class occurrence in the image to be classified.

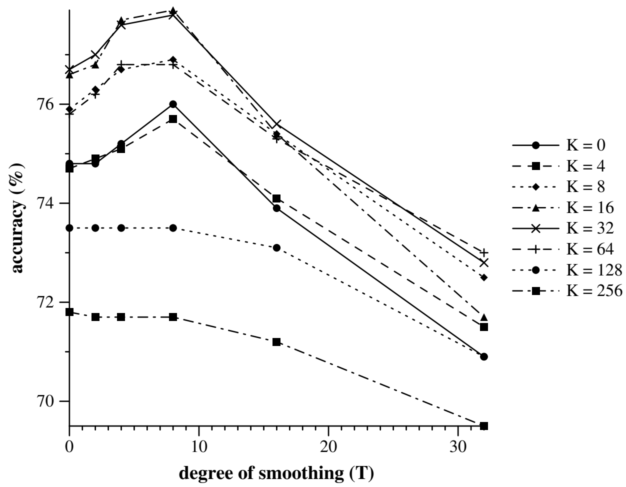
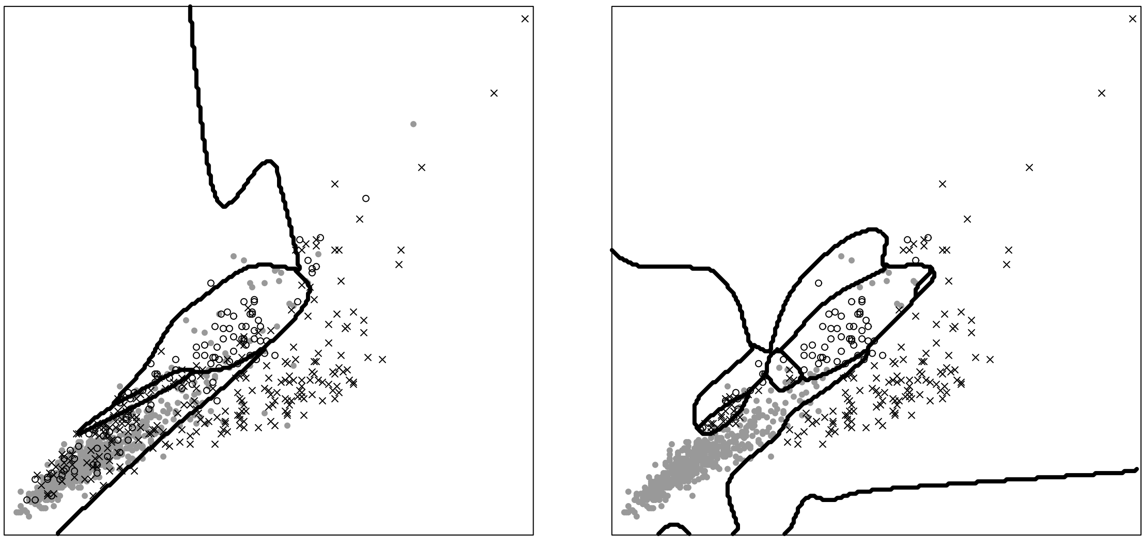


Figure 5: Glandorf anisotropic diffusion and k-filtering testcase

The second experiment tested the effect of spatial smoothing in the images. Since the Portugal data set did not contain ground truth data covering a spatially continuous area, we could not test the effects of spatial smoothing on this data set. Edge preserving anisotropic diffusion was applied separately to each band of the Glandorf image (see figure 5). If we ignore the two lowest plots in figure 5 which also show the effect of extreme filtering in feature space, we see that an improvement in accuracy can be achieved with a moderate amount of smoothing. For the Glandorf image, a value of T=8 as smoothing parameter for the diffusion algorithm consistently seems to give the best results, regardless of the k smoothing parameter used for feature space filtering. The second spatial smoothing method consisted of applying an iterative majority filtering to the classified output images. The effect of iterative majority filtering was tested on the Glandorf image. The original accuracy of 77.3% was improved to 80.0% after one iteration of the majority filtering algorithm. After 22 iterations the algorithm stabilised and no further changes were made to the image. Using multiple iterations as compared to using the algorithm in a non-iterative way however, only increased the general classification accuracy to 78.7%, an effect that can be attributed to over-smoothing. We will have another look at the effect of iterative majority filtering when discussing the combined spatial-spectral testcase.

The third series of experiments involved filtering of the training class clusters in feature space by means of k-nearest neighbour method. The effect of relabelling pixels with the majority label of their nearest neighbour is shown in the scatter plots of figure 6. Figure 6a shows the original cluster distribution for the red and blue bands of the Glandorf ground truth data. Figure 6b shows the resulting training set after k-nearest neighbour filtering applied to these two bands only. The confusion existing because of overlapping classes is clearly reduced by the nearest neighbour filtering. Reducing the confusion caused by overlap however, does not necessarily increase the correctness of the training set. The aim of this filtering is rather to decrease the difficulties encountered by classifiers such as neural networks to generate correct decision boundaries. Figure 6 also shows the change in decision boundaries caused by filtering the training data. The pixels in feature space that are detected by the



(a): Original training data (b): K-filtered training data with k=16

Figure 6: Glandorf training data : red band versus blue band

k-nearest neighbour method as being outliers can either be deleted or relabelled with the majority class label of their neighbours. Both methods were compared for all three testcases. It was found that the deletion method gave consistently better results than relabelling (see figures 3, 4 and 7). However, the overall effects of k-nearest neighbour filtering are not as clear-cut as the effects of the other experiments.

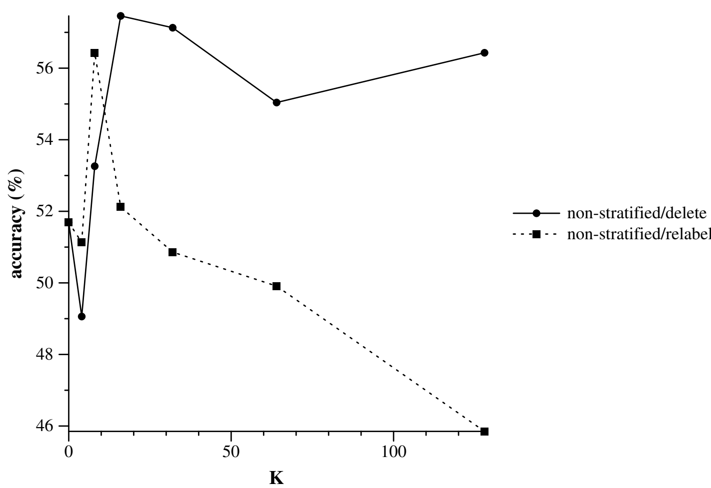


Figure 7: Portugal testcase

In the case of the Glandorf image (figure 3), for all the combinations of stratified/non-stratified and deletion/relabelling, as k is increased there is an initial slight decrease in accuracy followed by an increase to an optimum followed by a further decrease in accuracy. The graph for the Portugal testcase (figure 7) follows a similar pattern. The general behaviour of the CORINE testcase (figure 4) is similar, but the low and noisy accuracy values make these results the least reliable. This can be explained because the predefined classes for the CORINE map that we used as ground truth data are not necessarily related to spectrally distinguishable features. Since CORINE classes were designed with human photointerpreters in mind rather than automatic classification algorithms, this is not surprising.

The effect of k-nearest neighbour filtering in feature space is also shown in combination with different values for the spatial diffusion smoothing parameter (see figure 5). Again, the same general behaviour can be observed.

In the fourth set of experiments, we tried to include spatial context in the classification process by integrating spatial and feature space filtering according to the scheme depicted in figure 1. In the first part of these experiments, we combined anisotropic diffusion (spatial smoothing) and k-nearest neighbour filtering (spectral filtering). The results of these experiments are plotted in figure 5 and show that between certain bounds for the k value in the nearest neighbour filtering, spectral filtering improves the result obtained by spatial filtering alone, and vice versa. This implies that spatial and spectral coherence are not necessarily linked to each other. In a

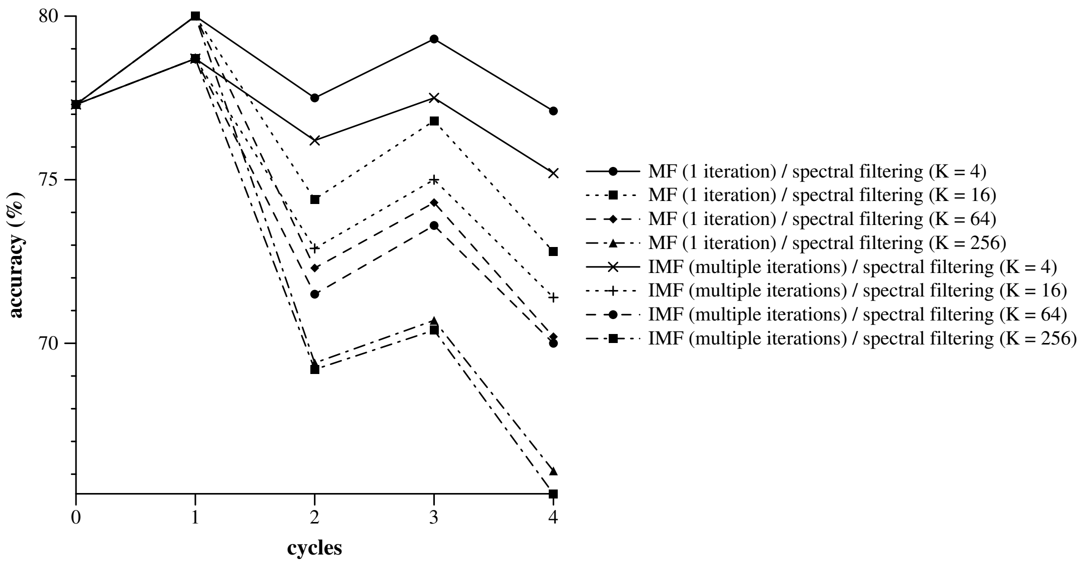


Figure 8: Glandorf spatial-spectral testcase

second set of spatial-spectral combination experiments we repeatedly applied iterative majority filtering (spatial) followed by k-nearest neighbour filtering (spectral). The results here however, are disappointing. In figure 8, the zero value on the abscissa shows the initial classified image at cycle zero, odd numbered cycles on the abscissa represent majority filtering, and even numbered cycles represent subsequent spectral filtering in feature space. Whereas the majority filtering shows a clear improvement in accuracy at each stage where it is applied, mapping the results of the majority filtered images back into feature space followed by k-filtering produces an overall loss in accuracy. Contrary to the previous experiments where spectral filtering in feature space was applied to the training data set prior to training the neural networks, k-filtering is now applied as a post-processing step to the spectral values of the classified image which are then mapped back into the classified image. As mentioned before, spatial and spectral coherence are not necessarily related. This can explain why the spectral filtering step decreases rather than increases the accuracy. Apparently the increase in spectral coherence happens at the expense of the spatial coherence in such a way that more is lost in terms of spatial coherence than is gained in terms of spectral coherence. Increasing the k-value in the spectral filtering seems to make matters worse. We only tested this combined method on the aerial Glandorf image, where each object is a relatively large and spatially coherent region. Therefore, it is not clear whether the conclusions can be extended to satellite images as well. Since completely covering ground truth for the satellite testcases was missing (Portugal) or too unreliable (CORINE), we could not repeat the experiment for these images.

In the last experiment we investigated the combined effects of network complexity and k-nearest neighbour filtering in feature space. For this purpose, we used a number of neural nets, each with one hidden layer, while varying the number of nodes in the hidden layer (see figure 9). We trained each network for five hundred cycles of the backpropagation algorithm. One problem of comparing different neural network architectures is that training every network with the same number of cycles discriminates against the more complex networks since these need

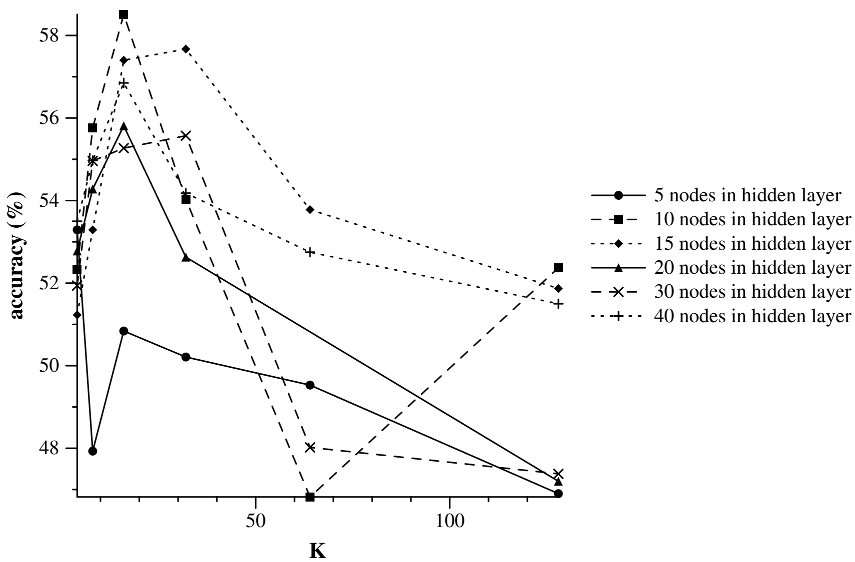


Figure 9: Networks of various complexities

more cycles in order to be trained properly. On the other hand, training the networks until the same global network error is reached would have favoured the simple networks as well since the more complex ones could probably do better. It is difficult therefore to identify the best network architecture from the graph. However, we can conclude that the k-nearest neighbour filtering has a similar effect on most architectures, namely an improvement followed by a degradation of accuracy as k is increased.

7 DISCUSSION

We have described various method for pre- and post-processing both training and image data in order to remove outliers and increase the accuracy of the classification. Whereas spatial processing of the data is a fairly standard technique, both for pre- and for postprocessing, we also experimented with filtering techniques that operate in feature space. Analogously to spatial filtering, feature space filtering has the effect of removing outliers in the data. In the feature space filtering approach however, these outliers are not defined by their spatial, but solely by the spectral context. We tested our methods on different data sets and our experiments verify that accuracy can be improved in all cases. Each method individually was capable of increasing the accuracy. In some cases further improvements could be gained by combining the methods. Not all effects shown in the graphs are equally easy to explain. In order to be able to draw conclusions that can be generalised over a wide range of remote sensing classification applications, further experiments are required to tell us whether these effects are statistically significant or just specific to our data sets.

8 REFERENCES

[1] J. A. Benediktsson, P. H. Swain, and O. K. Ersoy. Neural network approaches versus statistical methods in classification of multisource remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 28(4):540–552, 1990.

[2] D. L. Civco. Artificial neural networks for land-cover classification and mapping. *International Journal of Geographical Information Systems*, 7(3):173–186, 1993.

[3] R. G. Congalton. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing Environment*, 37:35–46, 1991.

[4] J. O. Eklundh, H. Yamamoto, and A. Rosenfeld. A relaxation method for multispectral pixel classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(1):72–75, 1980.

[5] P. E. Hart. The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, (May):515–516, 1968.

[6] G. F. Hepner. Artificial neural network classification using a minimal training set: Comparison to conventional supervised classification. *Photogrammetric Engineering and Remote Sensing*, 56(4):469–473, 1990.

[7] I. Kanellopoulos, A. Varfis, G. G. Wilkinson, and J. Mégier. Classification of remotely-sensed satellite images using multi-layer perceptron networks. In T. Kohonen, K. Makisara, O. Simula, and J. Kangas, editors, *Artificial Neural Networks*, pages 1067–1070, North-Holland, 1991. Elsevier Science Publishers B.V.

[8] M. Nagao and T. Matsuyama. Edge preserving smoothing. *Computer Graphics and Image Processing*, 9:394–407, 1979.

[9] Devijver P.A. and J.V. Kittler. *Pattern Recognition : a statistical approach*. Prentice-Hall, Englewood Cliffs, New Jersey, 1982.

[10] C. S. Penrod and T.J. Wagner. Another look at the edited nearest neighbor rule. *IEEE Transactions on Systems, Man and Cybernetics*, (February):92–94, 1977.

[11] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990.

[12] M. D. Richard and R. P. Lippmann. Neural network classifiers estimate Bayesian a posteriori probabilities. *Neural Computation*, 3:461–483, 1991.

[13] P. L. Rosin. Ellipse fitting by accumulating five-point fits. *Pattern Recognition Letters*, 14:661–669, 1993.

[14] G. G. Wilkinson. The generalisation of satellite-derived raster thematic maps for GIS input. *GIS*, 5:24–29, 1993.

[15] D. L. Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man and Cybernetics*, 2:408–420, 1972.