Improving Neural Network Generalisation

Paul L. Rosin and Freddy Fierens
Institute for Remote Sensing Applications
Joint Research Centre
I-21020 Ispra (VA), Italy
{paul.rosin,freddy.fierens}@jrc.it

Abstract — In this paper we study neural network overfitting on synthetically generated and real remote sensing data. The effect of overfitting is shown by: 1) visualising the shape of the decision boundaries in feature space during the learning process, and 2) by plotting the classification accuracy of independent test sets versus the number of training cycles. A solution to the overfitting problem is proposed that involves pre-processing the training data. The method relies on obtaining an increase of spectral coherence of individual training classes by applying k-nearest neighbour filtering. Points in feature space with class labels inconsistent with those of the majority of their neighbours are removed. This effectively simplifies the training data, and removes outliers and local inconsistencies. It is shown that using this approach can reduce the overfitting effect and increase the resulting classification accuracy.

OVERFITTING

Neural networks have become increasing popular over the last decade as an alternative to statistical approaches for classification of multispectral remote sensing data [2, 3]. Contrary to statistical classifiers, neural networks do not rely on an *a priori* model of data distributions. As a result they are often used as black box systems. However, there are various parameters that need to be chosen carefully in order to produce good results. Examples include network type, size and architecture, training step size, stop criterion, learning algorithms, and data representation. In this paper we restrict ourselves to multilayer perceptron networks with backpropagation as a supervised learning algorithm.

We focus on one particular problem with learning which is typical for neural networks: their *generalisation* capabilities. Generalisation is the ability to train with one data set and then successfully classify independent test sets. Although continued training will increase the training set accuracy, the danger exists that test set accuracy decreases after a certain point. This is attributable to *overfitting* since only individual training samples are available rather than the true underlying distribution. This can cause some distortion or displacement of the decision boundaries.

It has been shown that overfitting is related to the network capacity, which itself is determined by the number of training samples and the number of weights [1] and variance of the training set [8]. This suggests that appropriate network parameters can be selected on a theoretical basis. Unfortunately, there are several problems:

- It has been proposed that the practical capacity of neural networks is less than their theoretical capacity [5, 7].
- Weigend [10] showed that in practice small networks can overfit as well as large networks.
- Schaffer [9] considers the avoidance of overfitting as just the introduction of an application dependent bias. This implies that there is no general non-parametric method to eliminate overfitting in all cases. Instead, the degree of bias needs to be carefully tuned to the data, otherwise it may degrade performance.
- Overfitting is a local distortion of the decision boundaries, and there is no reason to assume that this occurs simultaneously in the entire feature space – it may occur in different areas at different times. This implies that in some areas in feature space training should be continued while in other areas overfitting has already occurred. Therefore some kind of local learning may be necessary, e.g. Bottou and Vapnik [4].

In order to get a clear idea of when and where overfitting occurs we first experimented on a synthetically generated data set. For this experiment we used two two-dimensional Gaussian distributions. Both distributions are symmetric and have the same variance, resulting in a theoretically optimal linear discrimination rule equidistant between the two distribution means. The means of the distributions are separated by four standard deviations, resulting in some overlap which might cause overfitting behaviour. The sample sizes of both distributions

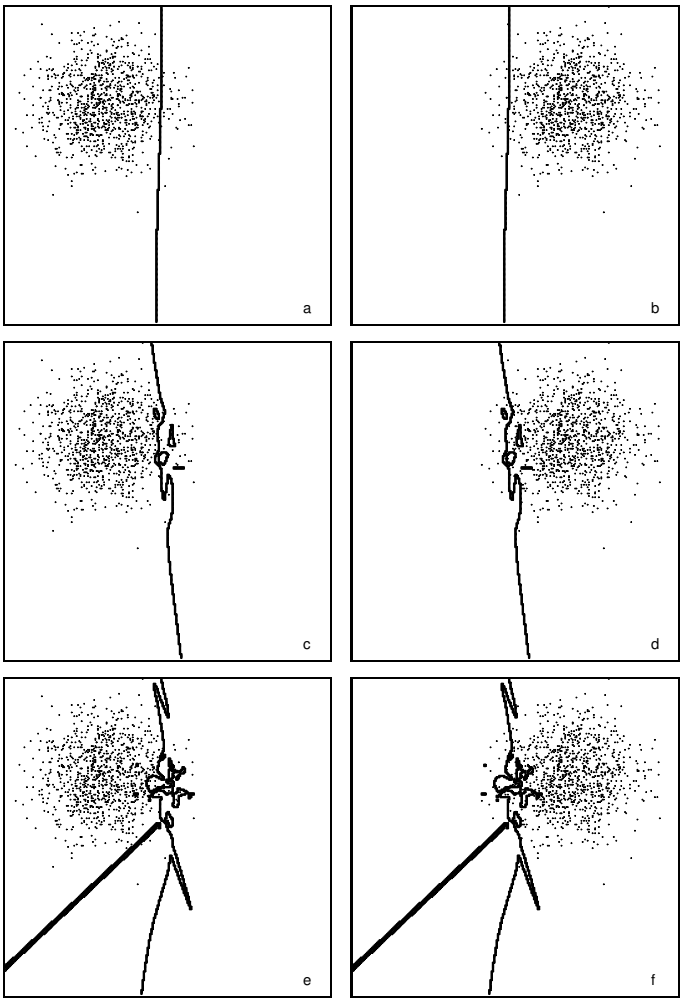

Figure 1: Feature space subdivision after 1 (a & b), 2000 (c & d), and 4000 (e & f) epochs showing overfitting. Figures a, c, e show training data from class 1. Figures b, d, f show training data from class 2.

are identical. This means that overfitting will not affect the overall position of the decision boundary but will only affect its shape. Fig. 1 shows three snapshots of the division of feature space during training, respectively after 1, 2000, and 4000 epochs of the backpropagation algorithm. The effect of overfitting is clearly visible in fig. 1c-f since the decision line deviates from the correct straight line, and small patches in the overlap region are incorrectly delineated resulting in incorrectly labelled islands in feature space.

EFFECTS OF FEATURE SPACE FILTERING ON OVERFITTING

We experimented with a data filtering technique in order to reduce sensitivity of the neural network learning process to outliers causing overfitting. We assume that sample pixels from a class exhibit spectral coherence. In other words, pixels from the same class should have similar spectral values, and therefore any pixels with very different spectral characteristics are outliers and are likely to cause confusion during the training phase. As shown in fig. 1 overtraining can force neural networks to construct local boundaries around outliers, rather than generalising by absorbing them in the surrounding class. In this experiment we investigate how increasing the spectral coherence of the feature space values decreases the likelihood that this problem occurs. We increase spectral coherence by deleting outliers in the training set. This is achieved by applying K-nearest neighbour filtering to the spectral values in the training set. K-nearest neighbour filtering compares each pixel in the training set to its K nearest neighbours defined in terms of Euclidean distances in feature space. If the class label of the central pixel is different from the majority of its K neighbours the pixel is considered an outlier and deleted from the training set. Deleting the pixel was shown to be more effective than relabelling it [6]. Increasing the value of the parameter K increases the degree of smoothing.

Overfitting behaviour is detected by plotting classification accuracy of the network during the iterative training process with respect to an independent test set. The test set was generated by taking another random sample from the two distributions. We hypothesise that overfitting occurs when this accuracy starts to decrease. At the same time the classification accuracy of the training data will generally continue to increase since it is the error on this training set that is being minimised by the backpropagation algorithm.

Each graph in fig. 2 shows the results of the neural network training and testing classification accuracy during the learning process applied to the synthetic data set described above. The different graphs represent different degrees of K-nearest neighbour filtering applied to the training data before initialising the learning process. Fig. 2a shows learning without data filtering, while figs. 2b-d show K set to 4, 16, and 64 respectively. All experiments made use of the same neural network architecture, consisting of 2 hidden layers with 15 nodes per hidden layer. Matching our hypothesis, the graphs show that the overfitting behaviour – a decrease in the test set accuracy during training – is reduced with increased filtering of the training data. Moreover, the final test set accuracies are im-

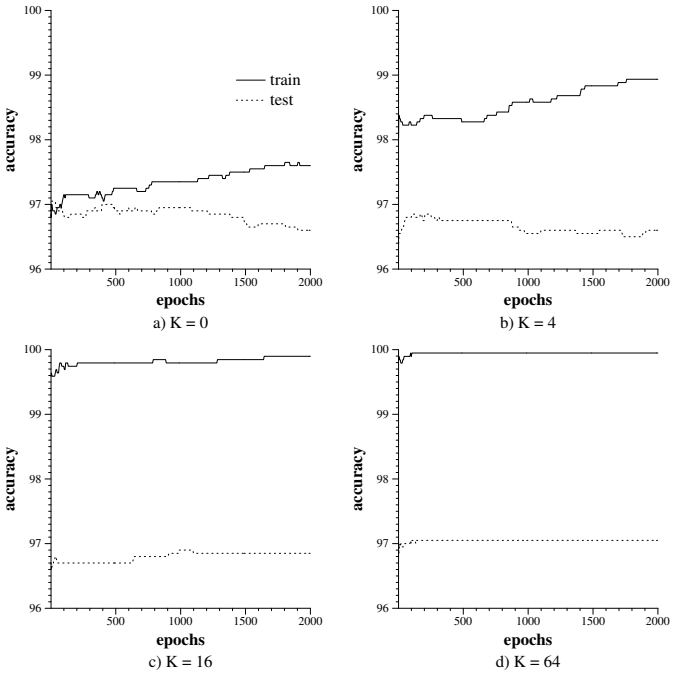

Figure 2: Overfitting as a function of test set accuracy for different degrees of feature space filtering

proved with increased filtering of the training data. Since the distributions of the two classes overlap, 100% test set accuracy is not possible, even with the optimal decision boundary (a vertical straight line equidistant from the two class means). Theoretically the peak of the test set accuracy curve is the optimal point to stop training. Because of the noisiness of this curve, however, it is difficult to use this as a reliable criterion.

The same experiments were also performed on a real data set taken from the 6 non-thermal bands of a Landsat-TM image of Portugal in 1991. This means that every data sample consists of six feature measures. The labels

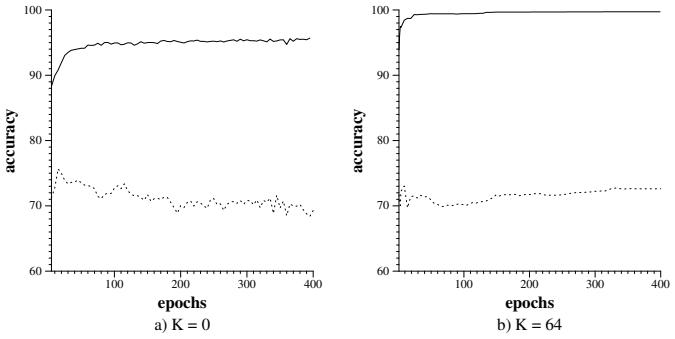

Figure 3: Overfitting and feature space filtering for a practical test case

in the training data were obtained from an on-site survey of the area and divided into 16 land cover classes. To make the experiment representative of real conditions the training and test data sets were obtained from separate areas. The training set contained 8047 samples, while the test set contained 4458 samples. The neural network architecture used in this case had 6 input nodes, two hidden layers with 20 nodes each, and 16 nodes in the output layer. Whereas in the synthetic example increased filtering of the training set consistently improved the test set accuracy, this is unlikely to occur in real situations where the optimal decision boundaries are more complex. Our experiments with a range of degrees of filtering show that for the Portugal data set the optimal amount of filtering is K=64. For the unfiltered training data, fig. 3a shows the typical overfitting behaviour where test set accuracy drops with increased training. In contrast, with K=64 filtering, fig. 3b shows that the test set accuracy improves with increased training. According to our hypothesis this is because overfitting is avoided.

REFERENCES

[1] E. B. Baum and D. Haussler. What size net gives valid generalization? *Neural Computation*, 1:151–160, 1989.

[2] 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.

[3] H. Bishof, W. Schneider, and A. J. Pinz. Multispectral classification of Landsat images using neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 30(3):482–490, 1992.

[4] L. Bottou and V. Vapnik. Local learning algorithms. *Neural Computation*, 4:888–900, 1992.

[5] R. P. W. Duin. Superlearning and neural network magic. *Pattern Recognition Letters*, 15:215–217, 1994.

[6] F. Fierens and P. L. Rosin. Filtering remote sensing data in the spatial and feature domains. In Jacky Desachy, editor, *Image and Signal Processing for Remote Sensing, Proc. SPIE vol. 2315*, pages 472–482, 1994.

[7] M. A. Kraaijveld and R. P. W. Duin. The effective capacity of multilayer feedforward network classifiers. *Proc. ICPR, Israel*, B:99–103, 1994.

[8] J. E. Moody. The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In J. E. Moody et al., editor, *Advances in NIPS-4*, pages 847–854. Morgan Kaufmann, 1992.

[9] C. Schaffer. Overfitting avoidance as bias. *Machine Learning*, 10:153–178, 1993.

[10] A. S. Weigend. On overfitting and the effective number of hidden units. In M. C. Mozer, P. Smolensky, D. S. Touretzky, J. L. Elman, and A. S. Weigend, editors, *Proceedings of the 1993 Connectionist Models Summer School*, pages 335–342, Hillsdale, NJ, 1994. Lawrence Erlbaum Associates.