A Neural Adaptive Algorithm for Feature Selection and Classification of High Dimensionality Data

Elisabetta Binaghi[1], Ignazio Gallo[1], Mirco Boschetti[2], and P. Alessandro Brivio[2]

[1] Dipartimento di Informatica e Comunicazione, Universita' degli Studi
dell'Insubria, Varese, Italy
{elisabetta.binaghi, gallo}@uninsubria.it
[2] CNR-IREA, Institute for Electromagnetic Sensing of the Environment,
Via Bassini 15, 20133 Milan, Italy
{boschetti.m, brivio.pa}@irea.cnr.it

Abstract. In this paper, we propose a novel method which involves neural adaptive techniques for identifying salient features and for classifying high dimensionality data. In particular a network pruning algorithm acting on Multi-Layer Perceptron topology is the foundation of the feature selection strategy. Feature selection is implemented within the back-propagation learning process and based on a measure of saliency derived from bell functions positioned between input and hidden layers and adaptively varied in shape and position during learning. Performances were evaluated experimentally within a Remote Sensing study, aimed to classify hyperspectral data. A comparison analysis was conducted with Support Vector Machine and conventional statistical and neural techniques. As seen in the experimental context, the adaptive neural classifier showed a competitive behavior with respect to the other classifiers considered; it performed a selection of the most relevant features and showed a robust behavior operating under minimal training and noisy situations.

1 Introduction

Recent applications of Pattern Recognition and in particular of Image Analysis and Classification deal with high dimensionality data.

In this context, the use of automated classification procedures is still limited by the lack of robust methods able to cope with the intrinsic complexity of high dimensionality and the consequent Hughes phenomenon, implying that the required number of labeled training samples for supervised classification increases as a function of dimensionality [1, 2]. The problem can be addressed in two complementary ways: -identify a classification model less sensitive to the Hughes phenomenon and/or -reduce the dimensionality of data and redundancies by applying feature selection strategies. Neural networks seems to be very good candidates for simultaneous feature selection and classification [3]. In view of these considerations, we designed an experimental study to investigate the robustness of a non conventional classification model when dealing with high dimensionality data. The model integrates feature selection and classification tasks in a unified framework based on adaptive techniques [4] built on the top of conventional Multi-Layer Perceptron [5].

F. Roli and S. Vitulano (Eds.): ICIAP 2005, LNCS 3617, pp. 753 – 760, 2005.

The model was experimentally evaluated within a Remote Sensing study aimed to classify MIVIS hyperspectral data. Robustness was evaluated in terms of performance under different training conditions and in the presence of redundant noisy bands. The model was compared with conventional statistical classifiers, Multi-Layer Perceptron and SVM.

2 Adaptive Neural Model for Feature Selection and Classification

The use of neural networks for feature extraction and selection seems promising since the ability to solve a task with a smaller number of features is evolved during training by integrating the process of learning with feature extraction (hidden neurons aggregate input features), feature selection and classification [6]˙.

This work presents a supervised adaptive classification model built on the top of Multi-Layer Perceptron, able to integrate in a unified framework feature selection and classification stages. The feature selection task is inserted within the training process and the evaluation of feature saliency is accomplished directly by the back-propagation learning algorithm that adaptively modifies special functions in shape and position on input layer in order to minimize training error. This mechanism directly accomplishes the feature selection task within the learning stage avoiding trial and error procedures which imply multiple training runs.

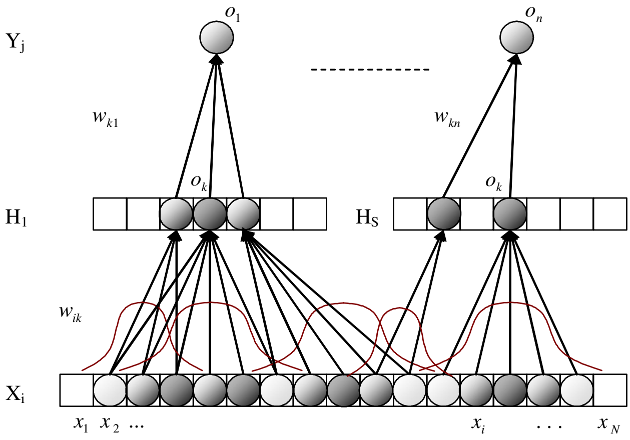


Fig. 1. Topology of the proposed adaptive Neural Model, characterized by one network for each class

Fig. 1 presents the topology of the adaptive model conceived as a composition of full connected neural networks, each of them devoted to selecting the best set of feature for discriminating one class from the others. The feature selection mechanism is embedded between input and hidden layers connections. Special functions (Figure 2a, 2b) are defined to modify connection weights: they act as penalty function for connection values and then weight the importance of features associated with the concerned input neurons.

The modified aggregation function I_k for adaptive neurons is described in the following formula

$$I_k = \sum_{i=1}^{M} w_{ik} \cdot o_i \cdot h_{ks}(i) \tag{1}$$

with M maximum number of input connections for the *j-th* neuron; o_i output of the *i-th* input neuron;

$$h_{ks}(i) = L_l(i; p, c_{ks}, a_{ks}) - L_r(i; p, c_{ks}, b_{ks}) = \frac{1}{1+e^{-p(i-(c_{ks}-a_{ks}))}} - \frac{1}{1+e^{-p(i-(c_{ks}+b_{ks}))}} \tag{2}$$

is the *s-th* bell function of the *k-th* hidden neuron; L_l and L_r are two sigmoid functions; p controls the slope of the two sigmoid functions; a_{ks} and b_{ks} controls the width of the bell function h_{ks} and c_{ks} is the centre of h_{ks}.

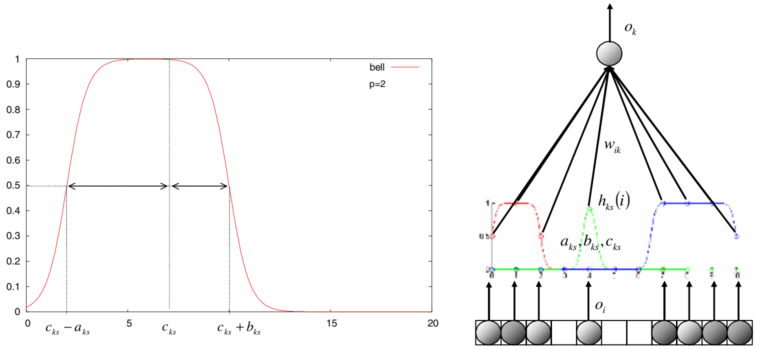


Fig. 2. (a) The bell function h_{ks} in Eq. 4 and (b) derived feature selection mechanism; $h_{ks}(i)$ represents the feature saliency measure for the *i-th* input feature

2.1 Network Configuration and Neural Learning

The neural learning procedure, aimed to identify discriminating classification functions, includes a non-conventional sub-goal formulated as the search for the most adequate number of bell functions h_{ks} varying adaptively in position and shape to lead to the smallest training error. The goal is achieved within the back-propagation learning scheme by applying the delta rule training algorithm [7] to standard weights w_{ik} and parameters a_{ks}, b_{ks} of bell functions.

At each learning step, the variation of parameters a_{ks}, b_{ks} results in a new positioning of the corresponding bell functions h_{ks}; the model attempts to re-defines the range of each bell function minimizing the overlap for all the bell functions associated with each hidden neuron; the maximum overlap allowed is at the inflection point of two adjacent bell functions.

2.1.1 Bell Function Removing and Insertion

For each bell function the distance between c_{ks} - a_{ks} and c_{ks} + b_{ks} is reduced at each learning step acting on a_{ks}, b_{ks} as follows:

$$a_{ks} = a_{ks} - r \cdot rnd \; ; \; b_{ks} = b_{ks} - r \cdot rnd \tag{3}$$

where r is a reduction rate with $0< r \leq 1$ and rnd is a random number with $0< rnd \leq 1$. The reduction mechanism is inserted within the overall learning process in such a way that back-propagation is able to compensate erroneous reductions.

Consequently bell functions are removed when all the following conditions are satisfied

1. $(c_{ks} + b_{ks}) - (c_{ks} - a_{ks}) < MIN\_WIDTH$, where MIN_WIDTH is a threshold value
2. $h_{ks}(m) < \dfrac{0.1}{1 + w_{mk}^2}$, where m indicates the neuron with connection weight having maximum value among those associated with connections under the bell function h_{ks}.

Variation of distance between c_{ks} - a_{ks} and c_{ks} + b_{ks} during learning can lead to a progressive increment of function areas which implies in general a decrease of connection significance. A bell function with a distance over the maximum allowed value and with mean connection weights $w_{ik} \cdot h_{ks}(i)$ under the threshold is split into two functions.

2.1.2 Removal of a Hidden Neuron

Feature selection with this type of neural net can lead to a progressive architecture simplification. In fact, as a consequence of the bell function removal mechanism, a hidden neuron can become useless for the classification task. This occurs when all the bell functions are removed by the procedure explained above, and this in turn happens when no significant connection exists between this hidden neuron and all input features.

This pruning mechanism is fundamental for training speed up and in many cases leads to a hidden layer with only the minimum number of neurons i.e. two.

An important aspect of this method is that we do not need to retrain the network after removal of a neuron and relative synapses, because the neuron was excluded by the learning procedure.

2.1.3 Initialization of the Neural Model

Initialization of the adaptive neural model involves specification of the following topological aspects:

- Number of bell functions for each neuron
- Number of neurons for each hidden layer

The proposed model is designed to cope with high dimensionality data. Considering that the number of bell functions can increase during learning by means of an insertion mechanism and that hidden neurons can be removed by the criteria stated above, we may pose a heuristic initialization criterion which defines the minimal initial number of bell functions equal to two for each hidden neuron.

The initial number of hidden neurons can be heuristically assessed according to conventional configuration rules [5] and specifically considering, the advantage of an automatic reduction of useless hidden neurons as a function of the input dimensions.

3 Experimental Evaluation

Our experiments were designed to assess the robustness of the adaptive neural model in classifying high dimensionality data . In particular empirical tests were conducted addressing the following main questions:

- how did the performances of the neural adaptive model depends upon different levels of supervised knowledge available for training ?
- how did the neural adaptive model compare with statistical and neural classifiers ?

Experiments were conducted within a remote sensing study aimed to classify MIVIS hyperspectral data. The study area represents a typical agro-ecosystem belonging to Ticino River regional park and located south west of Milan, Italy. A detailed land cover map of this area was obtained by integrating field surveys with aerial photo interpretation thus providing labeled data for the experiments.

The source data is constituted by an hyperspectral image with a total of 102 spectral bands acquired by MIVIS (Multispectral Infrared and Visible Imaging Spectrometer) with an aerial survey over the study area. Spectral bands were reduced to 51 by eliminating noisy bands and to 92 by eliminating thermal infrared range.

Sample areas for five classes (rice, corn, bare soil, poplar, natural forest) were chosen having good spatial coverage so that the natural variability of land cover class could be ensured. Three types of sets were chosen for training, named T1, T2 and T3 having different cardinality: 100, 52 and 25 pixels respectively.

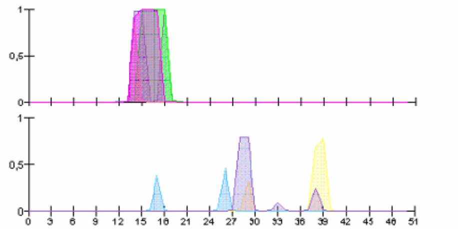
Test set was composed of 60 pixels for each cover class, randomly selected outside of training areas, by applying stratified technique, that guarantees a level of confidence, in the overall accuracy estimation, of 95% for all classes [8]

3.1 Robustness Evaluation Under Minimal Training Conditions

The experiment aimed to evaluate the performances of the proposed adaptive model when trained under three different conditions of pattern cardinality for each class.

In order to isolate factors related to training set cardinality, the overall training was facilitated by introducing a simple feature selection pre-processing stage aimed at eliminating noisy bands. The total number of features selected correspond to 51 spectral channels.

The adaptive model performed a selection of the most relevant features during training. Fig. 3 shows the bell functions assessment within the neural network topology after training with the T3 sample set, for corn and soil classes better emphasizing the feature selection mechanism. The figure exploits the situation for each of the two sub-networks devoted to the classification of a given class. The last column



Features selected	class name	number of hidden neurons	
		initial	final
	Mais	51	6
	Soil	51	3

Fig. 3. Graph shows bell function assessment after training with T3 sample; columns 3 and 4 show the number of neurons for corn and soil classes, before and after training procedure

shows how the adaptive process simplifies the topology of the network: the number of initial hidden neurons was 51 for each class, but after training task many neurons were removed.

Performances of the adaptive neural model were evaluated for all the training conditions and compared with those obtained from the Maximum Likelihood (ML), the Spectral Angle Mapper (SAM) [9], a specific hyperspectral classifier, (both implemented in the software ENVI [10]), the Multi-Layer Perceptron (MLP) and the Support Vector Machine (SVM) with Radial Basis Function kernel type; parameters c and gamma were tuned using the grid tool LIBSVM [11, 12].

The agreement between reference test data and classification results was analyzed by means of the confusion matrix. Fig. 4 shows the overall accuracy (OA) values obtained for all the classifiers considered when trained with the three data sets.

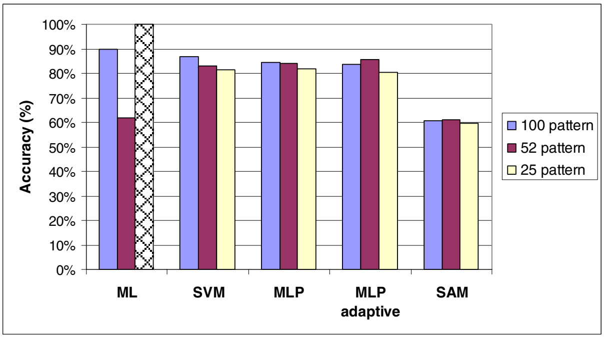


Fig. 4. Overall testing accuracy of the different methods in different pattern cardinality

The adaptive method presents a stable behavior under the three different training conditions reaching an high level of accuracy in all cases (over 80%). Maximum Likelihood is strongly influenced by the training conditions: performances are superior in case of training with T1, reaching an accuracy value close to 90% and drops to 60% using training set T2. Training T3 is not applicable. The SAM algorithm shows a stable behavior due to the fact that the classification is based on the calculation of the distance from an average spectrum per class, but the accuracy reached is in all cases inferior to that of the adaptive model. Conventional MLP and SVM show a stable behaviour and performances comparable with those obtained by our model.

3.2 Robustness Evaluation Under a Noisy Situation

This experiment aimed to evaluate robustness of the adaptive neural model in dealing with redundant and noisy data. To this purpose the classifier was trained with the T1 data set considering all the 92 bands available (i.e. excluding thermal infrared).

Basing on positions and values of the bell functions in the trained network, we deduced the results of the feature selection procedure.

Fig. 5 shows the features selected for rice class; feature selection obtained adaptively starting from 92 bands was compared with feature selection obtained starting with pre-selected 51 bands. Results are mostly consistent. Accuracy obtained by the adaptive neural model was evaluated and compared with those obtained by the ML, MLP, and the SVM (Table 1). All the classifier registered a decrease in OA passing from 51 to 92 bands in input; however the adaptive model and ML outperformed the other three; in addition the lowest decrease was registered for the adaptive model.

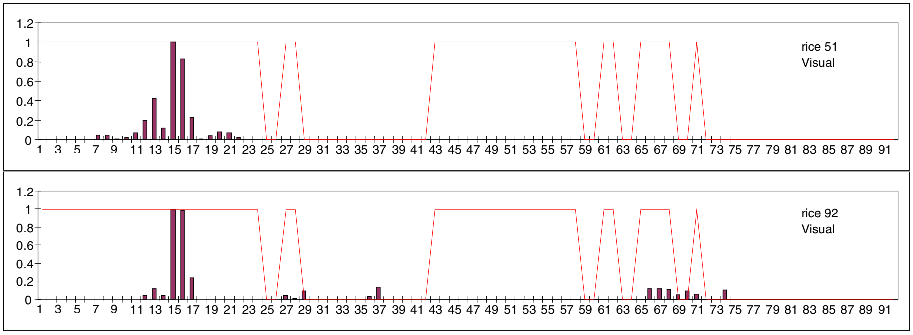


Fig. 5. Feature selection result for the rice class, compared with the preliminary band selection based on coefficient of variation

Table 1. Comparison of Adaptive MLP, ML, MLP, and SVM in terms of OA accuracy, when 51 pre-selected bands and overall 92 bands are presented in input

Number of features	51	92	
Number of training patterns	100	100	delta
ML	89.67%	76.00%	13.67%
MLP	84.67%	59.33%	25.34%
SVM	86.67%	72.00%	14.67%
AMLP	83.67%	75.33%	8.33%

4 Conclusion

In this paper, we have proposed the use of an adaptive neural network model for the twofold task of feature selection and classification. The feature selection mechanism is embedded within the back-propagation learning algorithm and directly accomplished during the learning process without implying multiple runs. Two critical aspects were investigated: minimal training and noisy situation. As seen in our experimental context, the features selection strategy allows proper selection of relevant features obtaining, as side effects, the reduction of the topological complexity of the model during training. Accuracy results obtained allow to conclude that our model can be considered an adequate tool in remote sensing studies when a feature selection phase is not possible or unadvisable and/or limited supervised data are available.

References

1. Fukunaga, K., Hayes, R.R., 1989. Effects of Sample Size Classifier Design, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 11, no.8, pp.873-885
2. Jain A.K., Duin R.P., Mao J., 2000. Statistical Pattern Recognition: a review, IEEE Trans. on Pattern Analysis and Machine Intelligence, 22, pp.4-37
3. Jain, A., Zongker D., 1997. Feature Selection: Evaluation, Application, and Small Sample Performance, IEEE Trans. on Pattern Analysis and Machine Intelligence, 19 (2), pp.153-158.
4. Pao, Y.H., 1989. Adaptive Pattern Recognition and Neural Networks. Addison Wesley, MA
5. Bishop, C.M., 1995. Neural Networks for Pattern Recognition, Oxford University Press, Oxford.
6. Reed, R., 1993. Pruning Algorithms - a survey. IEEE Trans. Neural Networks 5, 740-747.
7. Rumelhart, H., Hinton, G.E., Williams, R.J., 1986. Learning Internal Representation by Error Propagation, Parallel Distributed Processing, Rumelhart H., Mc Lelland J.L.(eds.), 318-362. MIT Press, Cambridge, MA
8. Van Genderen J.L., Lock, B.F., Vass, P.A., 1978. Remote Sensing: Statistical testing of thematic map accuracy, Remote Sensing of Environment, 7, pp. 3-14.
9. Kruse, F. A., Lefkoff, A. B., Boardman, J. B., Heidebrecht, K. B., Shapiro, A. T., Barloon, P. J., Goetz, A. F. H., 1993. The Spectral Image Processing System (SIPS) - Interactive Visualization and Analysis of Imaging spectrometer Data, Remote Sensing of Environment, 44, pp. 145 - 163.
10. ENVI, The Environment for Visualizing Images, Research Systems Inc., http://www.rsinc.com/envi
11. Vapnik V.N., Statistical Learning Theory. Wiley, New York, 1998.
12. Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.