UNIVERSITY OF JOENSUU
DEPARTMENT OF COMPUTER SCIENCE
Report Series A

The effect of irrelevant inputs on the generalization performance of MLP neural networks

Juha Hakkarainen
Report A-1998-7

ACM I.2.6
ISSN 0789-7316
ISBN 951-708-664-4

The effect of irrelevant inputs on the generalization performance of MLP neural networks

Juha Hakkarainen
Department of Computer Science
University of Joensuu

Abstract - *Irrelevant inputs decrease the performance of most induction based machine learning algorithms. In this paper, we study experimentally the effect of irrelevant inputs on neural network learning in two different training sample sizes by using two artificial and one real world datasets. The simulation results indicate that if enough training data is available, then the multilayer perceptron neural networks can be used in domains with high number of irrelevant inputs without significant loss in generalization capability.*

Keywords - feature selection, knowledge representation, neural networks

Introduction

The research of the irrelevant features[1] has gained a lot of attention in the machine learning community. Most of the work deals with the selection of relevant features (e.g. Battiti, 1994; Bradley *et al.*, 1996; Domingos, 1997). The feature selection methods can be divided into *wrapper* and *filter* methods (John *et al.*, 1994). The filter methods select the features independently of the learning algorithm while the wrapper methods use the learning algorithm itself to evaluate the relevance of the inputs. There exist numerous methods for feature selection in the literature; an overview can be found in Piramuthu (1996).

Different machine learning methods have different sensitivity to irrelevant inputs. Some simple learning methods, like the *nearest neighbour method,* ignore the issue of relevance entirely (Langley, 1994) and are therefore highly sensitive to irrelevant features. Some more advanced methods, like *multi-layer perceptron neural networks,* use weights to assign degrees of relevance to input features and

---

[1]. There exist different definitions for relevance and irrelevance in the literature (for review, see Kohavi and John, 1997). In this paper, we have used the word irrelevant as a synonym for uniformly random.

are therefore assumed to be more tolerant of irrelevant features.

Although many different feature selection methods have been proposed also for neural networks (e.g. Cherkauer and Shavlik, 1996; Battiti, 1994; Piramuthu, 1996) the actual effect of irrelevant inputs on the generalization performance of neural networks is little studied in the literature. We focus on this problem in this paper.

Experiments

To find out how the irrelevant features affect the generalization capability we trained and tested multi-layer perceptron neural networks with three different datasets and two different training sample sizes by adding random inputs to each dataset one by one. The following datasets were used:

1. *Sin dataset*. The inputs are points from 0 to 2Π and the targets are sin function values of the inputs. We created randomly two training sets (100 and 500 instances) and one test set (100 instances).

2. *Phoneme dataset* from ELENA project (Aviles-Cruz *et al*., 1995): This real world dataset consists of five attributes (the amplitudes of the five first harmonics of vowels) and two classes (nasal and oral). The original database consists of 5404 instances. We selected 1000 and 4000 instances for training and 1000 instances for testing.

3. *Concentric dataset*. This artificial dataset originally from ELENA project (Aviles-Cruz *et al*., 1995) consists of two classes: the points of class 0 are uniformly distributed into a circle of radius 0.3 centered on (0.5, 0.5) and the points of class 1 into a ring with radius 0.5 outside of class 0. There is no overlap between classes. The original database consists of 2500 instances. We used two training sets (1250 and 2000 instances) and one test set (500 instances).

We used multi-layer perceptron networks with one hidden layer in our tests. This type of networks have proved to be universal approximators - they are capable of approximating any Borel measurable function to any desired degree of accuracy (Hornik *et al*., 1989). The hyperbolic tangent

$$f(x) = \tanh(1.5x) \tag{1}$$

suggested by Kalman and Kwasny (1992) was used as an output function.

The networks used with the Sin, Phoneme and Concentric datasets had 4, 5 and 6 nodes in the hidden layer, respectively. To avoid problems related to premature saturation, the weights of the neurons in the hidden and output layers were initialized with random values inside the range ($-1/N$, $1/N$), where N is the total number of inputs of neuron. The networks were trained by using the *backpropagation* algorithm (Werbos, 1974; Rumelhart *et al*., 1986) with learning rate 0.001 and no momentum. The networks were trained 500 epochs with the Sin and Phoneme datasets, and 4000 epochs with the Concentric dataset.

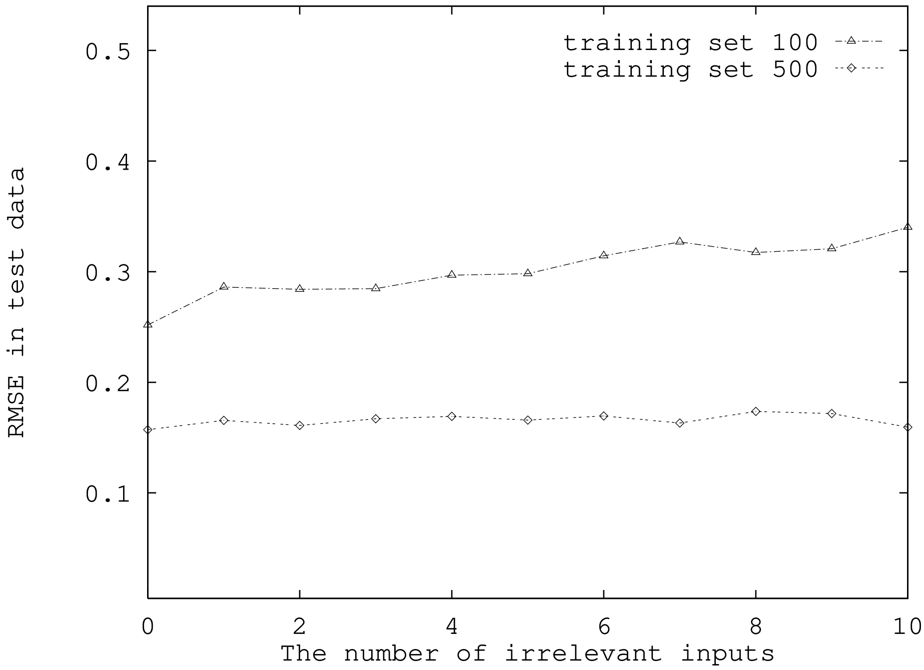

Figure 1. The RMSE error in the Sin dataset as a function of irrelevant inputs for training set sizes of 100 and 500 samples.

Results

The simulation results are presented in Table 1 and Figures 1, 2 and 3. Each data point presented is an average of 100 runs. The large standard deviations of some datapoints with Concentric dataset shown in Table 1 are caused by one or two networks which have not converged within 4000 epochs.

The results of the Sin and Phoneme datasets resemble each other. With smaller training set the irrelevant features have a noticeable effect on generalization, but with larger training set the irrelevant features increase the error in the test set only a little.

With the Concentric data, the difference between the results with larger and smaller training set is smaller than with two other datasets. This is somewhat expected since the difference of sizes between smaller and larger training set is also smaller. The irrelevant input features increase the generalization error, but the error percentage in the test set remains still under 3 even with 10 irrelevant features.

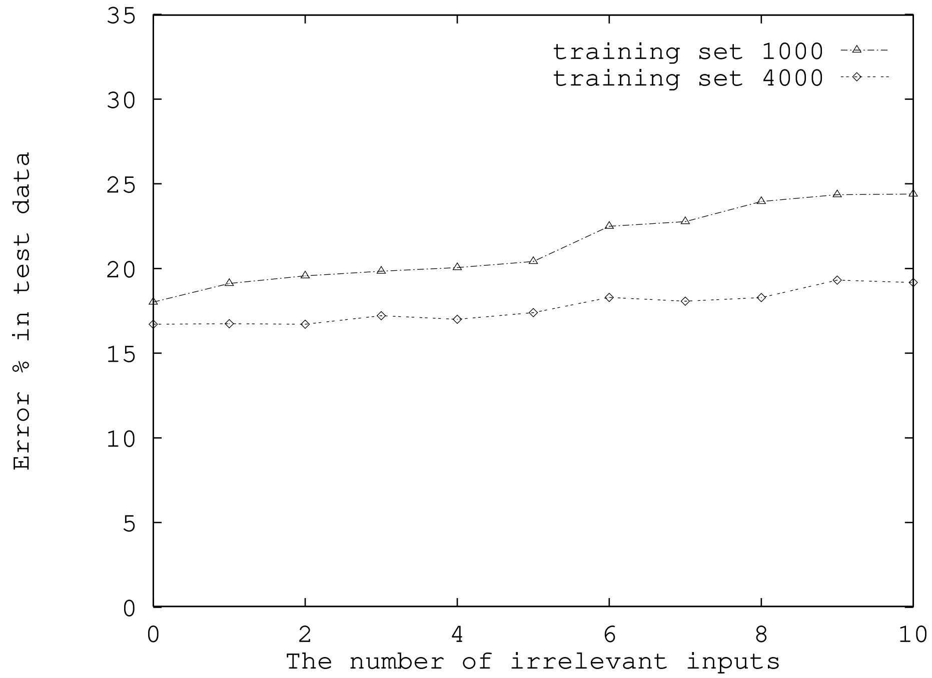

Figure 2. The error percentage in the Phoneme dataset as a function of irrelevant inputs for training set sizes of 1000 and 4000 samples.

Discussion

In this paper we have empirically tested the effect of irrelevant input features on the capability of multi-layer perceptron neural networks to generalize to new data. The simulation results indicate, that multi-layer perceptron neural networks can be used in domains with irrelevant features without significant loss in generalization performance. However, as the input dimensionality increases with irrelevant features, the size of the training set required for good generalization also increases (due to the curse of dimensionality).

If the amount of the available training data is small then the feature selection methods must be used also with neural networks. For example *evolutionary optimization* (Kyngäs and Hakkarainen, 1996) can be used. Since testing the goodness of the input representation is very time consuming by using neural networks itself, many filter methods like *decision tree algorithms* (Cherkauer *et al.* 1996), *mutual information based methods* (Battiti, 1994) and *stepwise multiple regression* (Piramuthu, 1996) have also been used for selecting features.

It should also be noted that, although the irrelevant features do not always disturb the capability of the networks to generalize to new data, they can increase the training time. They can also make the attempts to interpret the connection weights of neural network more difficult: it is possible that even if the neural network has a good generalization capability, some of the inputs are irrelevant.

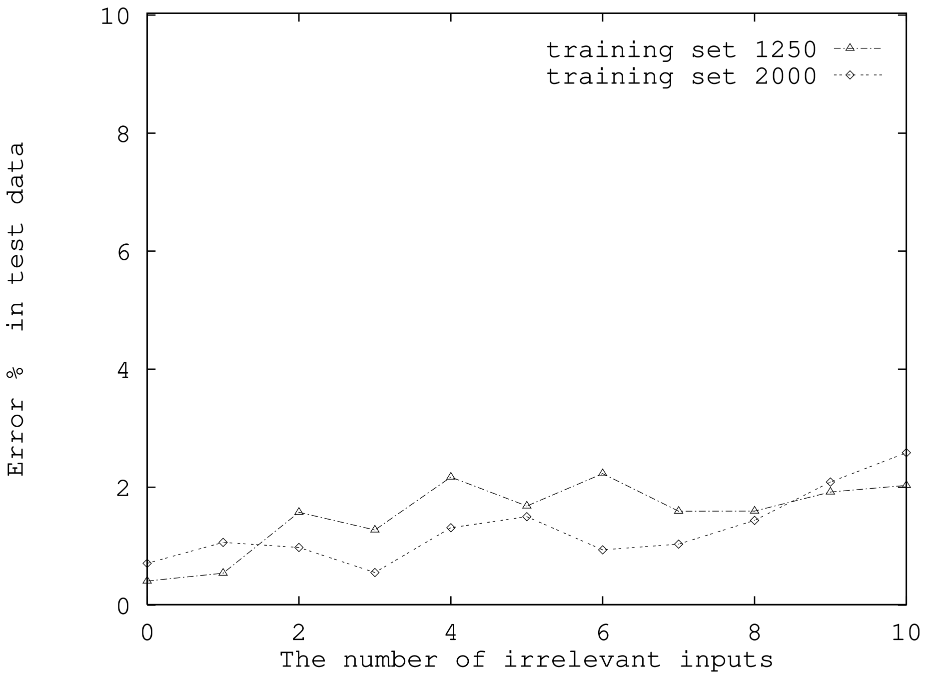

Figure 3. The error percentage in the Concentric dataset as a function of irrelevant inputs for training set sizes of 1250 and 2000 samples.

Table 1: Errors in test data (standard deviations are shown in parentheses)

The number of irrelevant inputs	Sin 100 RMSE	Sin 500 RMSE	Phoneme 1000 Error %	Phoneme 4000 Error %	Concentric 1250 Error %	Concentric 2000 Error %
0	0.252 (0.062)	0.157 (0.038)	18.014 (0.994)	16.707 (1.110)	0.410 (0.375)	0.710 (0.240)
1	0.286 (0.035)	0.166 (0.036)	19.119 (1.299)	16.742 (1.1585)	0.544 (0.243)	1.064 (0.262)
2	0.284 (0.032)	0.161 (0.024)	19.565 (0.941)	16.708 (1.129)	1.576 (4.926)	0.978 (0.195)
3	0.285 (0.033)	0.167 (0.030)	19.838 (1.359)	17.212 (1.365)	1.274 (0.260)	0.554 (0.324)
4	0.297 (0.032)	0.169 (0.026)	20.054 (1.060)	17.004 (1.114)	2.172 (4.353)	1.314 (0.362)
5	0.298 (0.036)	0.166 (0.027)	20.416 (1.101)	17.392 (1.534)	1.682 (0.455)	1.500 (0.590)
6	0.314 (0.044)	0.170 (0.033)	22.493 (1.679)	18.289 (1.734)	2.230 (0.479)	0.938 (0.192)
7	0.327 (0.034)	0.163 (0.029)	22.771 (1.571)	18.070 (1.635)	1.594 (0.365)	1.036 (0.353)
8	0.317 (0.041)	0.174 (0.031)	23.952 (1.586)	18.279 (1.622)	1.596 (0.331)	1.438 (0.549)
9	0.321 (0.036)	0.171 (0.028)	24.359 (1.432)	19.314 (1.620)	1.918 (0.484)	2.088 (2.774)
10	0.341 (0.045)	0.159 (0.037)	24.398 (1.637)	19.172 (1.533)	2.030 (0.320)	2.584 (0.882)

References

Aviles-Cruz C., Guérin-Dugué A., Voz J. L., & Van Cappel D. (1995). Deliverable R3-B1-P, Task B1: Databases. Technical Report ELENA-Nerves-II, ESPRIT-BRA No. B06891.

Battiti R. (1994). Using Mutual Information for Selecting Features in Supervised Neural Net Learning. *IEEE Transactions on Neural Networks*, 5, 537-550.

Bradley P. S., Mangasarian O. L., & Street W. N. (1996). Feature Selection via Mathematical Programming. Technical Report 95-21, University of Wisconsin.

Cherkauer K. J., & Shavlik J. W. (1996). Rapid Quality Estimation of Neural Network Input Representations. In, D. Touretzky, M. Mozer, and M. Hasselmo (Eds.), *Advances In Neural Information Processing Systems*, Vol. 8. Cambridge: MIT Press.

Domingos P. (1997). Context-Sensitive Feature Selection for Lazy Learners. *Artificial Intelligence Review*, 11, 227-253.

Hornik K., Stinchcombe M., & White H. (1990). Multilayer feedforward networks are universal approximators. *Neural Networks*, 3, 551-560.

John G. H., Kohavi R., & Pfleger P. (1994). Irrelevant features and the subset selection problem. *Proceedings of the Eleventh International Conference on Machine Learning*. New Brunswick: Morgan Kaufmann.

Kalman B. L., & Kwasny S. C. (1992). Why Tanh: Choosing a Sigmoidal Function. *Proceedings of the International Joint Conference on Neural Networks*. Baltimore: Morgan Kaufmann.

Kohavi R., & John G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97 (1-2), 273-324.

Kyngäs J., & Hakkarainen J. (1996). Predicting sunspot numbers with evolutionary optimized neural networks. *Proceedings of the Second Nordic Workshop on Genetic Algorithms and Their Applications*. Vaasa: University of Vaasa.

Langley P. (1994). Selection of Relevant Features in Machine Learning. *Proceedings of the AAAI Fall Symposium of Relevance*. New Orleans: AAAI Press.

Piramuthu S. (1996). The Effects of Feature Selection on Backpropagation in Feed-Forward Neural Networks. *Proceedings of HELNET International Workshop on Neural Networks (1994/1995)*. Amsterdam: VU University Press.

Rumelhart D. E., Hinton G. E., & Williams R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel Distributed Processing*. Cambridge: MIT Press.

Werbos P. J. (1974). Beyond regression: new tools for prediction and analysis in the behavioral sciences. Ph.D. thesis. Boston: Harvard University.