WAYS OF SELECTING INTERNAL PATTERNS IN MULTILAYER PERCEPTRON NETWORK

Marcin Kolibabka[1], Andrzej Cader[1],
Agnieszka Siwocha[1], Marcin Krupski[2]

[1] Information Technology Institute,
University of Social Science, Lodz, Poland
(mkolibabka, acader, asiwocha)@spoleczna.pl

[2] Department of Computer Science in Economics
Institute of Applied Economics and Informatics
Faculty of Economics and Sociology
University of Lodz, Poland
mkrupski@spoleczna.pl

Abstract

Creating and later learning one-way neural networks depends on many factors. Selecting many of them has estimated and experimental character. The suggested method is the Allows weakness of the influence of the not optimal choice of the net structure, also speed and momentum values are less influential in classic Back then Propagation Method. There are few modes of choosing elements to use in Followed algorithm

Key words: neural networks, artificial intelligence, back propagation

1 Introduction

Simple to implement one-way, multi-layer, non-linear neural networks called MLP (Multi-Layered Perceptron) [1] are conventional. For practical use of the network, however, it is necessary to construct an appropriate network structure as well as teaching it the proper reactions, relevant to the problem given.

The principles introduced in the late 80's of the twentieth century, describing the capabilities of neural networks - each limited continuous function can be approximated with arbitrarily small error by a network with one hidden layer [2,5], moreover, any function can be approximated with arbitrary accuracy by a network with two hidden layers [2,4] - and the development of algorithm of error back propagation (English EBP - Error Back Propagation) [4]

directly contributed to their prevalence, after earlier, long-term abandonment of research on them. For many applications, they are also predisposed by the relatively simple structure of the taught network, which combined with properly organized, parallel processing of signals allows for fast obtaining network's reaction to change of the input parameters.

Classification tasks are one of the key issues that are being solved with the usage of perceptron network. In the process of learning network „remembers" the patterns from the training ensemble and generalizes their forms in order to be able to recognize new input. This is obviously possible in the perfectly extending learning process. In practical applications such optimal solutions can be achieved by experimentation with learning parameters and network's structure. Each limitation of the number of experiments is therefore beneficial. Work developed method allows to reduce the impact of not-optimal network's structure and increases the speed parameter range of values and learning momentum, at which one achieves beneficial learning results. This method is an extension of the classic error back propagation method of enforcing a common standard for group of scales [6,7].

2 The selection of a multi-layer perceptron Network's structure

Selecting the proper number of layers and neurons for the usage of the network in the problem given has highly experimental nature. Kolmogorov's theorem for its theoretical nature has little practical significance, and even then it can only refer to a network with a single output and moreover with a linear activation function.

More significant in this regard is the statement:

Let's suppose that Φ is any continuous sigmoidal function. Then for every continuous function f defined in the [0.1] n> = 2, and for any ε> 0, there exists an integral number N and the ensemble of constants α_i, θ_i *and in j,i=1,..., N, j = 1, ..., n, such that the function*

$$F(x_1,...,x_n) = \sum_{i=1}^{N} \alpha_i \Phi\left(\sum_{j=1}^{n} w_{ij} x_j - \theta_i \right) \tag{1}$$

approximates the function f, ie,

$$|F(x_1, \dots, x_n) - f(x_1, \dots, x_n)| < \varepsilon \; for \; all \; \{x_1, \dots, x_n\} \epsilon [0,1]^n$$

However, it also has its limitations. For example, it cannot be used in classification problems for more than two groups.

Apart from the problem of selecting the number of layers, the proper selection of number of neurons in each of them has great importance. Obviously,

too small number of neurons prevents network from learning, because the network has too small information capacity then. Alternatively, one could select too big network, but this solution has even several disadvantages. The least troublesome is the extension of the learning time. The most impending the usage of the network is the fact that the redundant network tends to „over-learn". It manifests by a loss of the ability to generalize knowledge, which means that the network can recognize only the data from the training ensemble in such case. It cannot properly identify the data, which are in scope of the task domain, but have not been used during the learning phase.

The number of neurons hidden in the network allows to estimate the so-called Vapnik- Chervonenkis dimension (VCdim) [8]. This dimension for the ensemble of functions is defined as the maximum number of vectors, that can be grouped in all possible ways, by using the function from this ensemble. For the neural networks, it allows to estimate the generalization capabilities through expressing the relationship between them, the amount of learning samples, network's learning error and the generalization error. Unfortunately, the assignation of this dimension is usually very difficult and the evaluation is a very „imprecise".

$$2\left\lfloor \frac{K}{2} \right\rfloor N \le VC\dim \le 2N_W\left(1+\lg N_n\right) \tag{2}$$

where:
K - the number of neurons in the hidden layer
N - size of the input
N_W- the number of network scales
N_n- the number of networks neurons

In practice, this requires tedious testing networks with different amounts of neurons anyway. Such testing requires a cyclic learning, testing, and removing the redundant scales. And even using the algorithms: Optimal Brain Damage [9] and Optimal Brai n Surgeon to reduce the network's structure, does not accelerate the process of obtaining its optimal working significantly. Therefore, it would be beneficial to obtain such a learning process that would allow the network with not optimal structure, to work as well as the optimally structured network.

3 The method of enforcing the internal formulas

In methods from the error back propagation group the algorithm is based on the assumption of minimizing the error E. This value is the sum of the errors calculated for each training data vector. In such methods, a change in the learning scale value depends directly only on its previous value. None of

these changes is combined with the change of the other scale in the same iteration, and even in different iterations this change is indirect through the value of the inherited error.

In the method of enforcing internal formulas it has been proposed, in the learning process, adding additional relations between selected scales [7]. These relations can be very simple. In the simplest case, it is the sum of scales.

$$\sum_{w \in B} w = const. \tag{3}$$

Where B is the ensemble of selected for the „interlock" scales. Interlock word was used in quotation marks, because in reality one does not lock individual value of scales, and only their sum. So in the process of learning the different scales may change, however in the way that their sum remains constant - change then depends also on changes of other scale values in the group. Because of this in the solution space a hiperface is selected, on which the solution is being searched.

Adding the condition caused the necessity to modify the redundancy defining the change of a single scale. This condition can be taken into account by using the method of Lagrange'a multiplicators.

After the introduction of the condition [7] we get to solve the set of linear equations:

$$\begin{aligned} & w_{1,i+1}^{p} + w_{2,i+1}^{p} + \cdots + w_{n,i+1}^{p} = C \\ & w_{1,i+1}^{p} - \lambda = w_{1,i+1} \\ & w_{2,i+1}^{p} - \lambda = w_{2,i+1} \\ & \cdots \\ & w_{n,i+1}^{p} - \lambda = w_{n,i+1} \end{aligned} \tag{4}$$

Where $w_{1,i+1}^{p}, w_{2,i+1}^{p}, \ldots, w_{n,i+1}^{p}$ are searched values of the scales in step i+1, $w_{1,i+1}, w_{2,i+1}, \ldots, w_{n,i+1}$, scale values resulting from the classical method of error back propagation. The system can be easily solved, and the result gives new scale values.

Blocking the sum of the scales is not only possible to use redundancy between the scales. Another type of relation between the scales can be their product. In this case, however, to keep the flexibility and speed of resolution one should be reduced with blocking the scales up to pair in the ensemble B:

$$w_i^k * w_j^k = C_k \; where \; B_k = \{w_i^k, w_j^k\} \tag{5}$$

k - the number of another interlock and j k(w and k, in j l) is scale grouping in k-numbered relation. Therefore, in order to obtain new scale values in the next iteration it is necessary to solve ensemble k system of three equations. In every system there is one non-linear equation, the one with the interlock condition. On the other hand, systems of equations themselves are independent from one another, which is ensured by the divisibility of the ensembles B_k.

Interlock in the form of the sum allows for finding the minimum of hyper planes which are parallel to each other, while a multiplicative relationship changes the direction of the search, which may be advantageous for the targets set.

4 The ways of selecting the scales to „interlock" esembles

In the testing phase, various criteria of selecting the scales in the form of both the interlock of the sum and the products were examined. For the additive relations the first method was the selection of scales with the highest absolute value (maxAbs) [7], as the ones, that have quantitatively the greatest opportunity to influence the result of the networks performance. Resulting directly from the above method is the reverse method, which is the selection of scales with a value as close as possible to zero (minAbs).

The third and fourth method include scales, the change of values of which resulted in the largest and respectively the smallest, change of error on the result of networks performance in relation to the scales values (maxRatio [7] / minRatio). The tests used a threshold value δ by which the scales were changed, afterwards the full calculation was carried out for the learning data without modifying the values of scales. Obtained at the end root-mean-square error at the output of the network was divided by the value of the scale. One selected to interlock the scales, for which the so obtained value was the highest, or in the opposite method, the smallest.

Another method, not algorithmic anymore, was a manual selection of scales. It showed, that blocking all the scales in the neighboring neurons decreases the learning results.

For many „interlock" ensembles selection criteria were many groups related with one another, because in addition to the ranking of the groups algorithms of division of the scales for more ensembles were also required. On the other hand, it was necessary to examine into how many ensembles the selected scales can be divided, which resulted in the need for parameterization

of the scales finding including the number of groups algorithms. Selecting the groups process consisted of several steps:

a. ranking arrangement the relative scales relative to criteria for one interlock maxAbs / minAbs, maxRatio / minRatio
b. decision on the number of the groups and the number of scales in each group
c. the way of selecting scales for each group

The first step is analogical to the one with the selection of one group. The second determines the parameters with which we induce the third step algorithm. In the second point were tested:

- the number of groups equal to the number of classification groups
- two groups
- the number of interlock ensembles a grade larger than the number classification groups
- in the next stages of learning increasing or decreasing the number of groups

Having the scales sorted out and information about the number of groups constructed as a selection criteria of them into the respective ensembles. For every above mentioned case, the described experiments were checked for different amounts of scales in the interlock ensemble. Obviously in the presented method for product interlock, the number of scales in the group is stiffly set to two.

Having sorted out the weight and the information about the amount of groups a group selection criterion in the respective sets is constructed. These criteria may be analogical to those described earlier for the single interlock group, but a multiplicity of groups significantly broadens the possibilities of choice. And so the basic criterion maxRatio can be modified in a number of ways. The simplest way is to assign a certain amount of scales with the highest module to the first group, subsequent to the second and so on („main" assignment). Because of this the most important, regarding the determined criterion, scales are grouped together.

Another applied solution was assigning the scales to every group in order, first with the highest module to the first group, the next one to the second and so on. Thus, for example, with four groups into each group there will belong scales from every fourth position from ordered by selected criterion structure („proportionate" assignment).

Another modification was such the selection of the groups, that within a single interlock ensemble were the scales, which values of the applied criterion are in balance. This means that the scales were paired up, the one with the biggest and smallest value of the module („equilibrium" assignment). Analogously, one can select the scales according to other criteria.

5 The results of tests on exemplary classifying networks

The tests on the method have been conducted on the networks with the optimized structure for the task given, as well as on the redundant structure. In the first case, the difference between the non-interlock method, and the learning with enforcing the standard was not significant. However, at the stage of searching for the optimal networks structure, learning with enforcing the standard, noticeably improved the efficiency of learning. Networks with to big structure have decreased ability to classify data, absent in the learning ensemble, so they generalize problem worse [6]. Blocking changes this situation.

In both algorithms with blocking the sum and the product, learning effectiveness was highly dependent on the selection of the number of the groups, the scales in the group and the method of their selection. The tests were conducted in the following way: the network structure was being generated, which was afterwards taught with the standard method, and the same network with the same initial scales with enforcing the standard and with the same parameters, such as learning speed and momentum, but with different methods of enforcing the standard. The criterion of improvement was the number of identified samples from the test ensemble. Tests were conducted on the problem of classification of irises and the „Zoo" classification (based on 16 features the animals were divided into 7 groups), the classification of the glass (10 features, the classification into two groups). The test files contain respectively 45, 30 and 24 samples. Experiments for each set of blocked scales were repeated several times and the results provided in this thesis are the average of several tests for each of the networks.

For all the three classification questions blocking the set containing more than 90% of scales from the network grouped in one ensemble resulted in the network almost completely ceased to learn (Table 1). No effect of the interlock was observed with blocking about half of the scales from the network with all the possible methods. The network was learning comparatively to the absence of the interlock (Table 2).

Table 1. Number of well examined samples at the average for 15 learning attempts of redundant network with different speed and momentum parameters for different methods with the interlock of the majority of the scales in the network

	Irises	**Zoo**	**Glass**
Classic BP with momentum	41.3	22.4	20.7
maxAbs	13.8	11.4	10.2
maxRatio	12.2	18.2	9.8

Table 2. Number of well examined samples at the average for 15 learning attempts of redundant network with different speed and momentum parameters, for different methods with the interlock of the half of the scales in the network

	Irises	**Zoo**	**Glass**
Classic BP with momentum	41.3	22.4	20.7
maxAbs	41.1	21.8	21.6
maxRatio	41.2	22.8	20.8

Another attempt was based on increasing the amount of blocked scales, every 5000 iterations. With maxAbs approach no improvement of the network was observed. However, with maxRatio and classifying network for the problem of „Zoo" smaller influence of the learning speed selection on the networks performance has been observed. As far as with the classical method the number of identified samples ranged from 10 to 25 depending on the selected learning speed n and the momentum than with the same parameters for the enforcing the standard method the interval was from 20 to 25

This effect was observed as well in case when in the first step half of the scales in the network were blocked by selecting them using the maxRatio method, and in subsequent steps, their amount was reduced to half. This time the benefit was observed in all three questions. The best results were obtained by blocking half of the scales in the first learning cycle (selected by maxRatio method), and in the next steps the ensemble was reduced by removing half of the scales (Table 3). The achieved results were on average 8.7% better than the conventional method.

Table 3. The number of correctly identified samples on average for 35 samples of learning of the redundant network with different speed and momentum parameters, for different methods with decreasing number of scales blocked in the following stages

	Irises Stage I	**Irises Stage II**	**Irises Stage III**	**Zoo Stage I**	**Zoo Stage II**	**Zoo Stage III**	**Glass Stage I**	**Glass Stage II**	**Glass Stage III**
Classic BP with momentum	20.1	35.7	41.6	11.7	18.4	22.9	13.8	19.5	21.9
maxAbs	19.4	33.1	41.1	10.4	18.4	21.8	12.4	19.6	21.6
maxRatio	20.9	35.8	41.9	15.8	19.1	26.2	12.1	21.0	22.7

The next stage of the study was the selection of more than one group. The studies of this criterion for the sum had to be linked to choosing of the method of selecting the scales for the inter locks ensemble.

It proved that with moving the scales using maxAbs method and learning with many groups the criterion of scales selection has little importance. In the extreme case the completely randomly selected scales to three groups in the problem of irises achieved the results compatible with the best result from the selection using algorithms. The mean values of results of the experiments with maxAbs tables scheduling are shown in Table 4.

Table 4. The number of well-recognized average samples for 10 experiments of learning of the redundant network with different speed and momentum parameters, for different methods with blocking different numbers of groups of scales and max-Abs criterion together with sum blocking

	Irises	**Zoo**	**Glass**
Classic BP with momentum	40.2	23.1	21.4
maxAbs 2 gro-ups	39.1	23.4	22.1
the number of groups as the number of net-work outputs	40.2	22.7	20.1
Number of groups 10 or more	40.1	39.4	21.5

In case of selecting for interlocking the sum more than one group and ma-xRatio scheduling criterion in some ways of scales selection to the ensemble the results did not differ significantly from the selection of one group with the exception of extreme cases, where on one hand the total discrepancy of the network occurred followed for the problem of glass classification and almost perfect performance of the network for the classification of irises. In the worst case for the glass classification the network in 4 cases in the 15 experiments did not give any correct classification. In the best experiment with the classi-fication of irises the achieved result was equal to the performance of the op-timal network. The overview of the results is shown in Table 5.

Table 5. The number of well-defined at the average samples for 15 to learn a redundant network with different speed and momentum parameters, for different methods with blocking of different amounts of groups of scales with maxRatio criterion as well as interlock of the sum

	main assignment	**equilibrium assignment**	**proportional assignment**
Irises BP	41.2		
Irises 2 Groups	40.1	40.9	40.5
Irises 3 Groups	40.0	44.1	41.5
Irises 30 Groups	39.2	42.9	40.5
Zoo BP	22.9		
Zoo 2 Groups	21.6	21.5	23.8
Zoo 7 Groups	21.1	23.2	21.4
Zoo 70 Groups	18.9	22.4	21.5
Glass BP	21.1		
Glass 2 Groups	12.1	21.5	20.2
Glass 10 Groups	19.6	14.2	22.3

6 Summary

The enforcing the internal standards method, regardless of its form can improve the redundant networks chances in recognizing the data, which is not part of the learning subject. Although the network continues to achieve worse results than the optimal network, but they are very similar and in situations where the quick reaction of the neutron network is needed without the seeking of the optimal structure the described method can bring measurable time benefits.

References

1. Tadeusiewicz R., 1993, *Sieci neuronowe*, Akademicka Oficyna Wydawnicza RM, Warszawa.
2. Cybenko G., 1989, *Approximation by Superpositions of a Sigmoidal Function*, Mathematics of Control, Signals, and Systems, Vol. 2 ,pp. 303–314.
3. Rutkowska D., Piliński M., Rutkowski L., 1997, *Sieci neuronowe, algorytmy genetyczne i systemy rozmyte*, Wydawnictwo Naukowe PWN, Warszawa.
4. Rumelhart D., Hinton G., Williams R., 1986, *Learning Internal Representations by Error Propagation*. Parallel Distributed Processing, Vol.1, pp.318–362.
5. Hornik K., Stinchcombe M., White H., 1989, Multilayer feedforward networks are universal approximators. Neural Networks, 2, pp. 359–366.
6. Rutkowski L., 2005, *Metody i techniki sztucznej inteligencji*, Wydawnictwo Naukowe PWN, Warszawa.
7. Kolibabka M., Cader A., 2006, Metoda wymuszania wewnętrznych wzorców w jednokierunkowej sieci klasyfikującej, Automatyka, 10, 3, pp. 497–502.
8. Haykin S., 1994, *Neural networks: A Comprehensive Foundation*, Macmikkan College Publishing Company, New York.
9. Bishop Ch.M., 1995, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, New York.