Regularization of neural networks for improved load forecasting in the power system

Stanis³aw Osowski[♣♠] and Krzysztof Siwek[♣]

[♣]Warsaw University of Technology, 00-661 Warsaw, pl. Politechniki 1

[♠]Military University of Technology, 00-908 Warsaw, ul. Kaliskiego 2, POLAND

Indexing terms: short-term load forecasting, artificial neural networks, regularization methods

Abstract

The paper presents the regularization procedure for the neural network reduction to obtain the best results of load forecasting in the power system. The OBD pruning method will be applied in the solution. The numerical experiments have been concentrated on the prognosis of the load in the power system. Two kinds of experiments will be described: 24-hour forecast and the forecast of the daily mean of the load. It will be shown that application of the regularization of the neural network employed for prediction, will result in significant improvement of the forecasting accuracy.

1. Introduction

The learning process of feedforward neural networks aims at the minimization of certain cost function built on the basis of the training data. It is believed that at the proper selection of the learning samples and at the appropriate learning, the trained network will behave well on the other data, not belonging to the training set. This is known as the generalization ability of the neural networks. The concept of generalization is not easy and all theorems concerning it rely on the statistical properties of networks [1,5].

The general rule, following from the statistical considerations, is to apply the smallest possible network structure that is able to fit the learning data [1]. This is the reason of regularization for the neural network. The regularization is understood here as the reduction of complexity of the network, that is designing the network structure of the smallest possible architecture. There are many reduction techniques, such as the penalty methods, first order sensitivity techniques and second order sensitivity methods [1,3,4,5]. Among many existing regularization procedures the most efficient are the second order sensitivity methods, like OBD and OBS [3,4,5].

The time series prediction corresponding to the power demand forecasting in the power systems belongs to the difficult prediction problems, requiring the application of large size neural networks. In such networks the problem of good generalization is rather serious and needs special attention.. To solve the problem of good generalization, in this paper we have applied the second order sensitivity criteria belonging to the class of OBD. The modification of the sensitivity formula will be proposed for obtaining the Hessian matrix. The numerical results concerning the prediction of the load consumption of the power system will be also presented and discussed. It will be shown that application of the regularization procedure results in significant improvement of the load forecasting accuracy.

2. Generalization properties of neural networks

The generalization of the neural network is a measure of how well the network performs on the data not seen at the learning stage, once the training is complete. Generalization is mostly influenced by three parameters: the number of representative samples used in the training phase, the complexity of the problem under consideration and the network size. Generally the bigger the network size the larger number of samples should be used in learning.

However choosing the proper network size is not easy. If the network is too small, it will not be capable of forming good model of the problem being solved. On the other hand, if the network is too big, it may be able to implement numerous solutions that are consistent with the learning data, but the general tendency is to memorize the data instead of learning the statistical features of the process, represented by these data. We would like to find a network whose size best matches the capability of the network to the structure of underlying problem, represented by the data. The solution to the problem of choosing the optimal size of the network belongs to the task of generalization theory of neural networks.

According to this theory [1,5] the most important is the notion of so called Vapnik-Chervonenkis dimension (VCdim), defined as the size of the largest set S of data samples for which the system can implement all possible 2^S dichotomies on S. The exact value of VCdim for MLP is unknown and we rely on the upper and lower bounds of it. In the case of multilayer perceptron as a rule of thumb we can use double number of weights as a rough estimation of true VCdim [5]. It has been found that to obtain good generalization properties of the MLP network we have to use the number of training samples much larger than VCdim.

However it is generally true that there is a large amount of redundant information contained in the weights of fully connected network. Thus it seems reasonable to eliminate some weights from the network and at the same time retain the functional capability needed to solve the problem. The process of cutting the weights is called pruning or regularization. The regularization of network has many advantages. First it reduces the complexity of the network and makes the learning process easier. Secondly at the same number of training

samples the reduction of weights leads to the improvement of the generalization ability of neural networks [1,5].

3. OBD pruning procedure

The best way for pruning the network is the application of sensitivity measures. From all used sensitivity measures the most effective are the measures based on the information contained in the Hessian matrix [1,3]. We apply here LeCun approach [3], based on the estimation of the so called "saliency" of weights. The saliency coefficient is determined on the basis of only diagonal terms of the Hessian matrix (the diagonal terms are dominant, since the Hessian is positive definite). This saliency of the weight w_{ij} is defined in the following way

$$S_{ij} = \frac{\partial^2 E}{\partial w_{ij}^2} w_{ij}^2 \qquad (1)$$

where E is the error function, defined usually in the Euclidean metrics. The weights with the smallest saliency values should be deleted from the network.

Fig. 1 presents the typical changes of the saliency values of different weights of the neural network at the application of the OBD procedure, arranged in an increasing order. The upper curve corresponds to the saliency and the lower one to the changes of the error function. It is seen that both curves are rather flat in most regions and the decision, what weights should be pruned is not easy, since the differences of saliency values are not high and it is difficult to define the proper level of cutting them. After close analysis of the curves it is seen, that the line of cut should be placed in the region of sudden change of saliency (in this case it may be the level of 38 or 46 neurons indicated by the vertical dashed line). Moreover it is evident, that the pruning should be done iteratively. It means that we have to train the network to a reasonable error level, compute the coefficients S_{ij} for all weights, delete some smallest saliency weights, resume training and repeat this procedure several times.

The important point in this procedure is the determination of the inverse of the Hessian. The calculation of the second order derivative matrix is rather time consuming. However in our

solution we have used the approximated Hessian matrix offered by the variable metric BFGS learning procedure of the neural network [2]. According to this procedure the updating of the weight vector **w** in (k+1) iteration is performed in the following way

$$\mathbf{w}_{k+1} = \mathbf{w}_k - h\mathbf{V}\mathbf{g} \qquad (2)$$

where h is the learning constant, **V** – the inverse of Hessian matrix and **g** – the gradient vector.

Since we train the neural network using variable metric method, where this Hessian matrix is available on line, we use it in our implementation of OBD. According to the BFGS strategy, updating the inverse of Hessian $\mathbf{V}=\mathbf{H}^{-1}$ is done iteratively at the learning phase according to the following scheme [2]

$$\mathbf{V}_k = \mathbf{V}_{k-1} + \left[\mathbf{1} + \frac{\mathbf{r}_k^T \mathbf{V}_{k-1}\mathbf{r}_k}{\mathbf{s}_k^T \mathbf{r}_k}\right]\frac{\mathbf{s}_k \mathbf{s}_k^T}{\mathbf{s}_k^T \mathbf{r}_k} - \frac{\mathbf{s}_k \mathbf{r}_k^T \mathbf{V}_{k-1} + \mathbf{V}_{k-1}\mathbf{r}_k \mathbf{s}_k^T}{\mathbf{s}_k^T \mathbf{r}_k} \qquad (3)$$

The variables $\mathbf{s}_k$ and $\mathbf{r}_k$ used in the equation denote the increments of the weight vector **w** and the gradient vector **g**, respectively in two subsequent iterations, $\mathbf{s}_k = \mathbf{w}_k - \mathbf{w}_{k-1}$ and $\mathbf{r}_k = \mathbf{g}_k - \mathbf{g}_{k-1}$. The starting value of **V** equals the unity matrix.

4. Regularization of the neural network for 24-hour load prediction

The prediction of 24-hour load pattern for the next day using multilayer perceptron is the most widely used neural approach to short term load prediction [6-11]. This application makes use of the universal approximation ability of the MLP network [1] to represent the generally unknown function, mapping the past loads of the system into the present forecasted load at *d*th day and *h*th hour. Our general model of the load in the power system is presented here in the form

$$P(d,h) = f\left(\mathbf{w}, r, u, P(d-1,h), ..., P(d-D, h-H)\right) \qquad (4)$$

where **w** represents the weight vector of the network, *H* and *D* - the number of past hours and days, respectively, influencing the prediction process, *r* - the type of the day (workday or holiday) and *u* - the season of the year (autumn, winter, spring and summer).

The neural network architecture (Fig. 2) associated with this mathematical model possesses certain number of input nodes (one or two to code the type of the day and the season of the year and some nodes to represent the loads of some past days) and 24 output linear neurons equal to the number of hours of prediction (24 hours ahead).

The number of hidden layers and neurons of sigmoidal non-linearity is subject to adjustment in an experimental way by training different structures and choosing one of the smallest one, still satisfying the learning conditions. On the basis of some numerical simulations we have found that optimal number of input nodes in our case is 23, which correspond to D=3 and H=4. The number of hidden layers was equal two and each layer contains 20 and 15 sigmoidal neurons, respectively. In this way the structure of the MLP network is described as 23-20-15-24, which corresponds to 1179 synaptic weights.

This is an quasi optimal network structure with respect to the number of hidden layers and number of neurons in each layer. The problem of network regularization is quite serious since the structure possesses more than 1000 synaptic connections. Many of these connections are not essential for the $\mathbf{x} \rightarrow \mathbf{y}$ mapping of the data and should be removed to increase the generalization ability of the network.

The network has been trained on the half of the available data of Polish Power System from the years 1992 – 1996 and then tested on all of them. The data of 1997 have been used only for testing. The application of the OBD pruning procedure has resulted in elimination of 179 weights out of 1179. It means more than 15% reduction of the number of weights. Testing the original and reduced network on the same data has revealed the overall improvement of prediction accuracy on the data not taking part in learning. Table 1 presents the details of this improvement in the form of the mean absolute percentage errors (MAPE).

The regularization of the network had week influence on the results concerning the data partly used in learning. The most impressive is the reduction of error for the year 1997, used only in testing mode. In this case the reduction of network complexity was very important and resulted in almost 17% decrease of the mean error.

However the most impressive is the reduction of maximum (peak) errors observed at different years under prediction (Table 2).

These errors appear from time to time as a result of some unpredictable events, like the sudden change of weather, unexpected event in the country, etc. The MAX errors have been reduced for all years and this reduction is quite high, changing their values from year to year.

Fig. 3 presents the relative improvement of MAPE (Fig. 3a) and MAX (Fig. 3b) errors calculated as the ratio of the change of corresponding error to the value of it before OBD regularization.

Observe that the change of MAPE errors observed for the learning data (1992-1996) are rather low. The real increase of accuracy is observed for the testing data, not taking part in learning (the year 1997). This is due to the generalization properties and their relationship to the complexity of the neural network.

Quite different results have been observed for MAX errors. These errors have been reduced in visible way (up to 45% in relative terms) for all years, of the data used both in learning and testing (1992-1996) and in testing only (1997). However the increase of accuracy for the data used only in testing and not taking part in learning, is one of the highest and exceeds 35%.

5. Regularization of the neural network for daily mean load prediction

The effect of regularization of neural network is mostly seen for large scale neural networks, like these used for 24-hour load prediction. However even in small size problems it may be quite evident. In this section we will show it on the example of the neural network used for forecasting the mean load consumption for the particular day of the year (the network of only one output neuron).

One of the most efficient approaches to this problem is once again the application of the multilayer perceptron. In our model we take the input vector to the network containing 23 input nodes, 13 hidden neurons and one output linear neuron. The input nodes represent the mean

loads of the days of the previous years, the season of the year (spring, summer, autumn, winter) and type of the day (workday or a holiday). The number of hidden neurons has been chosen on the ground of good generalization ability requirements of the network. The signal of output neuron represents the prediction of the daily mean of the load for the particular day. The appropriate structure of the proposed feed-forward neural network for prediction of the mean load is shown in Fig. 4.

The network has been trained and tested on the available data of Polish Power System from the years 1992 - 1997. Again the data of 1997 have been used only for testing. After application of the OBD procedure 65 weights (out of 326) of the network have been pruned. Fig. 5 presents the results in the form of simplified structures of the network. As it is seen not only individual weights have been cut, but also one hidden neuron has been reduced as a result of cutting its connection with the output neuron.

The highest decrease of the MAPE error of the mean load prediction has been observed for the year 1997, used only in the testing mode. The results of testing the network before and after regularization are presented in Table 3.

This time the reductions of both MAPE and MAX errors were not as significant as it was in the previous case. This follows from the fact that the initial number of weights of MLP was much smaller in this case. The MLP architecture has been carefully selected, leading to “almost optimal” one. However even in this case we have got the highest reduction of MAPE error for the year 1997, used only in the testing mode. It was almost 5% relative improvement of MAPE and 3% relative improvement of MAX errors.

6. Discussions and conclusions

The paper has presented the method of improvement of generalization ability of neural network by pruning the weights. Looking at good results of regularization of neural networks we should be aware that application of neural networks for load forecasting in not new and the MLP approach is not the only one. There are many papers [6-11] reporting the results of application

of different types of neural networks for 24-hour load forecasting. However, fair comparison of these results is not possible, since they have been obtained for countries of different size, incomparable climate conditions and different levels of industrial development. The amount of available data, especially weather conditions are also not equal.

For example the best results of load prediction for Spanish Power System reported in [7] obtained by using MLP were 3.40% (1992) and 2.93% (1993) of MAPE errors. For USA [9] the best reported results are 2.05% for weekdays and 2.47% for weekends (chosen months of the year) or 1.92% (the paper [10]). For Greece the application of fuzzy neural networks has allowed to reduce the MAPE error to the level of 1.67% [8].The best results reported for German system (Bayern region) correspond to 1.61% of MAPE error [11] and have been obtained by using specialized modular technique with the weather conditions correction. The results of load prognosis for England and Wales [6], reported only for weekends and holidays were all above 8.93% of MAPE error.

It is evident, that the results depend on the year under investigation, country, the applied method and are difficult to compare. However it is proved fact, that application of the regularization of the neural network leads to significant improvement of the accuracy of neural predictor. It means that our proposed regularization approach may find also application to improve the results of forecasting in any reported above methods.

References

[1] S. HAYKIN, 'Neural networks, a comprehensive foundation', (Macmillan, New York, 1994)

[2] S. OSOWSKI, M. STODOLSKI, P. BOJARCZAK, 'Fast second order learning algorithm for feedforward MLP', *Neural Networks*,1996, **9**, (9), pp. 1583-1597.

[3] Y. LeCUN, J. DENKER, S. SOLLA, 'Optimal brain damage', in "Advances of NIPS2", D. Touretzky, 1990, pp. 598-605.

[4] R. REED, 'Pruning algorithms – a survey', *IEEE Tr. Neural Networks*, 1993**, 4**, (3), pp. 740-747.

[5] D. HUSH, B. G. HORNE, 'Progress in supervised neural networks', *IEEE Signal Processing Magazine*, 1993, **10**, (5), pp. 8-39

[6] J. WANG, B. RAMSAY, 'A neural network based estimator for electricity spot-pricing', *Neurocomputing*, 1998, **23**, (1), pp. 47-57

[7] F. J. MARIN, F. SANDOVAL, 'Short-term peak load forecasting, statistical methods versus artificial neural networks', Proc. Int. Workshop on Artificial Neural Networks, IWANN'97, 1997, Malaga, Spain, pp. 1334-1343

[8] S. E. PAPADAKIS, J. B. THEOCHARIS, S. J. KIARTZIS, A. G. BAKIRTZIS, 'A novel approach to short-term load forecasting using fuzzy neural networks', *IEEE Tran. Power Systems*, 1998, **13**, (2), pp. 480-491

[9] I. DREZGA, S. RAHMAN, 'Short-term load forecasting with local ANN predictors', *IEEE Tran. Power Systems*, 1999, **14**, (3), pp. 844-850

[10] H. YOO, R. L. PIMMEL, 'Short-term load forecasting using self-supervised adaptive neural network', *IEEE Tran. Power Systems*, 1999, **14**, (2), pp. 779-784

[11] T. BAUMANN, E. THURNER, D. DEISENHOFER, 'Real world implementation of a short-term load forecasting method using modular neural networks', 12th Power System Computation Conf., PSCC'12, Dresden, 1996, pp. 231-237

Table 1 The summary of the MAPE errors of the load prediction in Polish Power System for the years 1992-1997

Year	MAPE of original network	MAPE of reduced network
1992	3.54%	3.53%
1993	3.48%	3.49%
1994	3.23%	3.22%
1995	3.11%	3.10%
1996	3.10%	3.00%
1997	3.61%	2.99%

Table 2 The summary of the MAX percentage errors of the load prediction in Polish Power System for the years 1992-1997

Year	MAX error of original network	MAX error of reduced network
1992	47.8%	27.0%
1993	44.4%	28.8%
1994	33.6%	27.1%
1995	32.1%	28.2%
1996	32.0%	26.5%
1997	47.2%	29.1%

Table 3 The summary of the MAPE and MAX percentage errors of the mean daily load prediction in Polish Power System for the years 1992-1997

Year	MAPE of original network	MAPE of reduced network	MAX error of original network	MAX error of reduced network
1992	1.62%	1.60%	6.08%	6.10%
1993	1.88%	1.87%	9.63%	9.60%
1994	1.97%	1.94%	14.64%	11.24%
1995	1.94%	1.94%	11.38%	11.12%
1996	1.77%	1.77%	12.14%	12.03%
1997	1.70%	1.62%	8.56%	8.32%

Captions of the illustrations

Fig. 1 The characteristics of saliency value of different weights of neural network in OBD procedure (the weights arranged in an increasing order)

Fig. 2 The MLP network for 24-hour ahead load prediction

Fig. 3 The relative change of prediction errors for the MLP network before and after OBD regularization: a) MAPE , b) maximum percentage errors

Fig. 4 The neural network for prediction of the daily mean of the load

Fig. 5 The structures of the neural network resulting from pruning of the weights

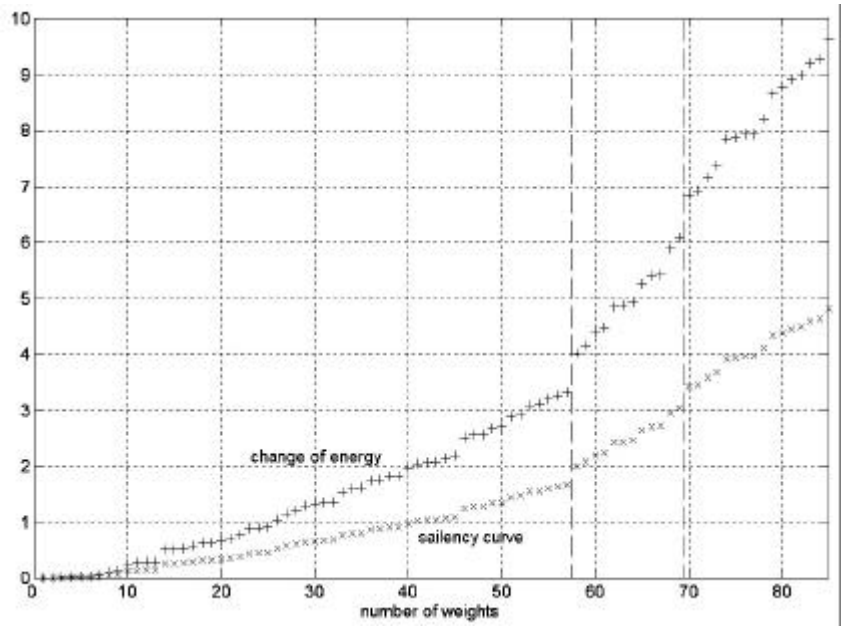
change of energy
saliency curve
number of weights
0
1
2
3
4
5
6
7
8
9
10
0
10
20
30
40
50
60
70
80

Fig. 1

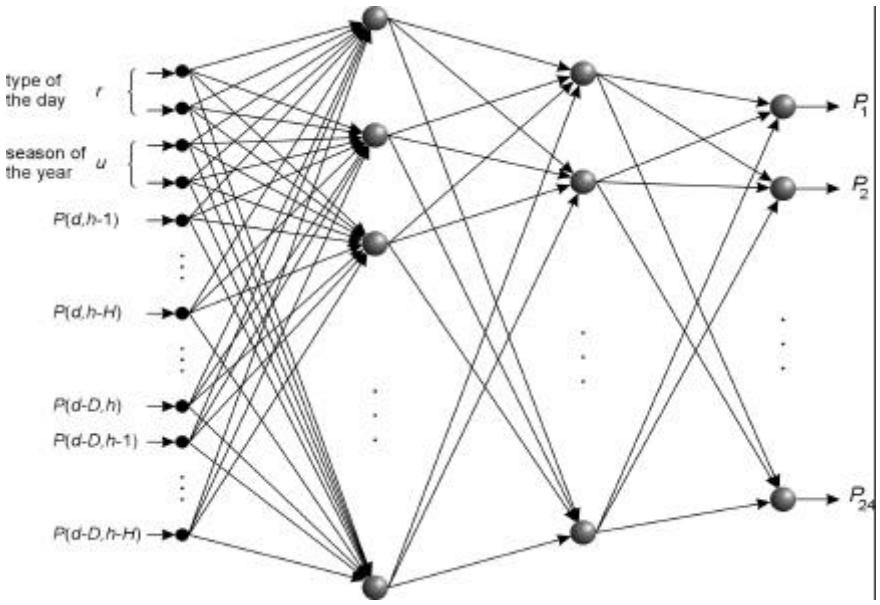
type of the day
r
season of the year
u
P(d,h-1)
P(d,h-H)
P(d-D,h)
P(d-D,h-1)
P(d-D,h-H)
P1
P2
P24

Fig. 2

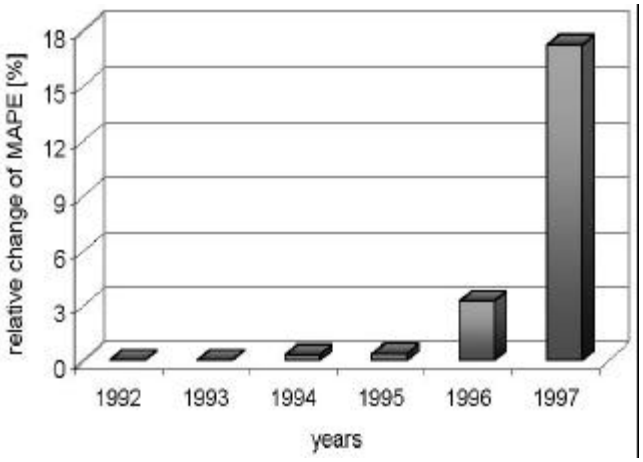
relative change of MAPE [%]
18
15
12
9
6
3
0
1992
1993
1994
1995
1996
1997
years

Fig. 3a

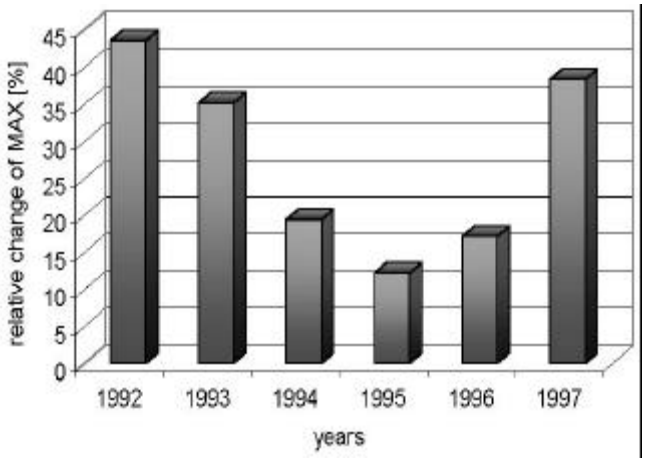
relative change of MAX [%]
45
40
35
30
25
20
15
10
5
0
1992
1993
1994
1995
1996
1997
years

Fig. 3b

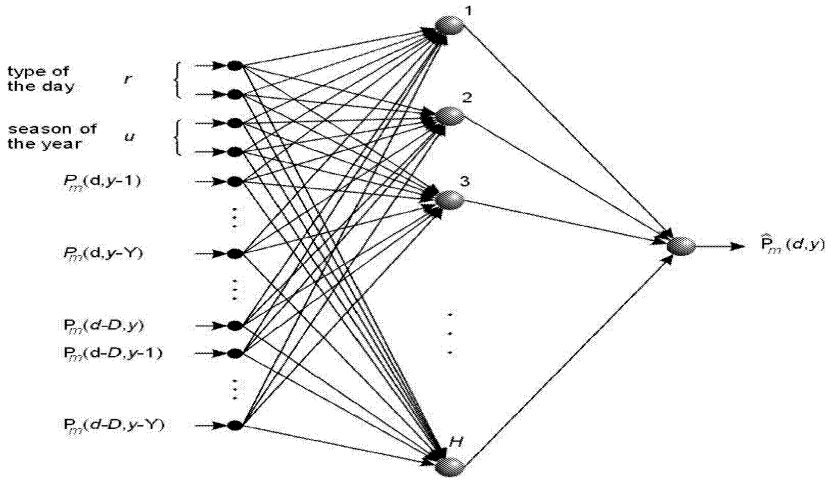
type of the day
r
season of the year
u
P_m(d,y-1)
P_m(d,y-Y)
P_m(d-D,y)
P_m(d-D,y-1)
P_m(d-D,y-Y)
1
2
3
H
$\hat{P}_m$(d,y)

Fig. 4

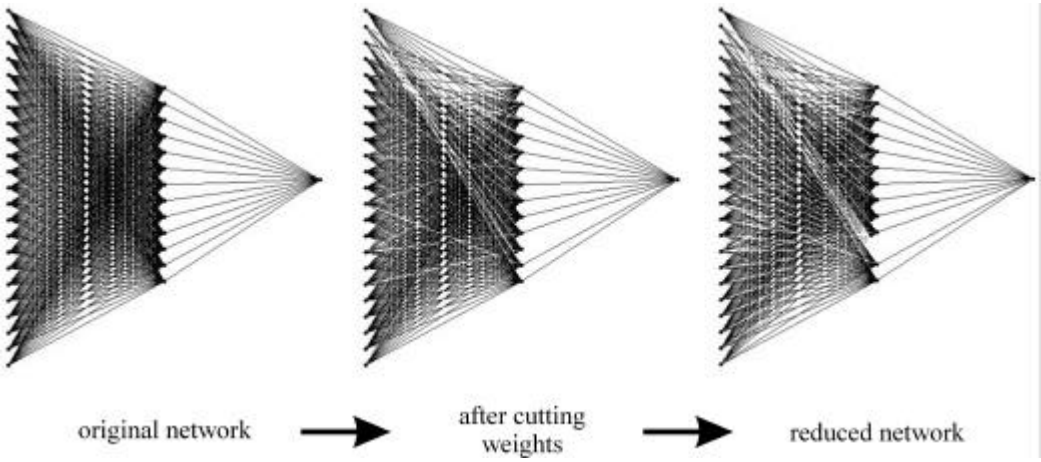

Fig. 5