Committee of networks

Ararat Vardanyan, Ashot Chilingaryan

Yerevan Physics Institute, Cosmic Ray Division, Armenia

Abstract:

The reduction of the errors committed by neural estimator has two aspects. The first one is the minimization of training error and the second minimization of generalization error. In present paper the "committee of networks" approach is presented in attempt to obtain a better generalization proceeding from a number of trained networks. Different types of committees are described and the experimental results are discussed.

Introduction:

It is common practice in the application of neural information technologies to train many different candidate networks and then to select the best, (on the basis of performance on an training data set) and to use further only this network. There are three disadvantages with such an approach. First of all, if we select the best network among many trained all of the efforts involved in training the remaining networks are wasted. Second, the network with best performance on training data set can be over-trained, and third, even if the over-training effects are controlled, the generalization error (of particular network from all available) on the test data might not be the smallest, if the training error is smallest, due to the sampling effects.

Different networks trained on the same training data set but with different training scenarios contain different information on the general nature of considered phenomenon, combination of these networks can give more generalized representation of these data, then the single "best" network.

Usually this procedure is called a committee of networks. In simplest case, when all networks are equally contributed, i.e. the final output is the average of all outputs of different networks, it is called a mean committee. Another type of optimal nonlinear combination of networks is discussed in the next section.

The committee procedures:

There are many different possibilities to make a committee of networks. In this paper the averaging and the median committees are presented. The first procedure is a simple averaging of all neural networks outputs and taking this mean as a final estimate. In ideal case, when the error of each trained network $e_i(x)$ has zero mean (unbiased training) and all errors of different networks are un-correlated:

$$e_i(x) = True(x) - Out_i(x) \; x \; is\, a\, training\, vector, \quad i = 1, \ldots N_{networks}$$

it is shown that even the simplest averaging error: $E_{com} = E_{av} / L$ where E_{com} is the committee error, E_{av} - averaged error of all networks and L is the total number of different networks (Bishop C.M.). In practice, the reduction of estimation error is much smaller because the errors are highly correlated. However, it is easy to show that the committee process cannot produce an increase of the expected error:

$$E_{com} \leq E_{av}$$

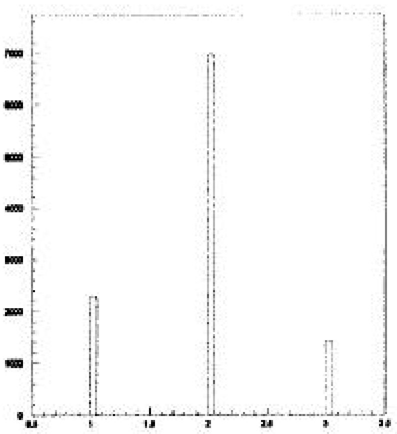

Figure 1: Contributions of different networks in final estimate (analytic models)

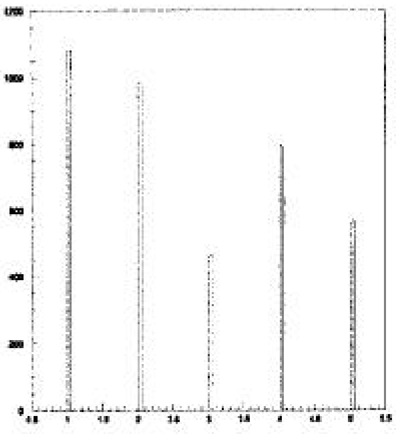

Figure 2: Contributions of different networks in final estimate

We might expect that some members of the average committee will make better predictions than other members. We would therefore expect to be able to reduce the error still further if we give greater weight to some committee members than to others. But the better prediction on training sample does not mean better prediction on control sample, so we do not know to which members to give greater weights. To automate the optimal weighting of neural networks we use the ordered statistics technique and call that procedure as a median committee, in analogy to the non-parametric estimate of multivariate probability density function, introduced in [2] (A. Chilingaryan 1989).

This procedure can be explained as follows: trained networks are implemented for the independent (control) data set and then the output values of all different networks are ordered in increasing (or decreasing) sequence and the middle (median) value of this sequence is taken as an final estimate. In this case again only one network's output is selected, but for each particular event as a median will be selected more stable reliable network from all available networks. So, all networks are implemented in estimation process, but each of them as many times as its output becomes the median of variation row of all outputs (fig. 1,2).

$$W_i = \begin{cases} 1, \text{ if } i = (L+1)/2, \text{ (L is an odd number)} \\ 0, \text{ otherwise} \end{cases} \quad (1)$$

If L is an even number, two middle outputs are taken with equal weights. The median committee looks somehow like weighted mean committee, because each network has a different contribution to the generalized final output. The problem how to give the weights to networks is avoided, because the predictions are weighted according to the ordered statistics.

In this case the weight of each network is defined in ordering procedure, using its output on control sample, which seems to be quite natural way to determine the weights. And the expectation of error reduction can be viewed as arising from reduced variance of network outputs for each individual event due to the removing large and small outputs (placed in the beginning and end of the variation row). Thus, the main attention in results discussing will be paid to the median committee procedure.

Resolving of the mixture of analytic models:

For the demonstration of the advantage of committee estimator comparing with the single best network the combined regression function is used, which consist of three different analytic models. The regression function in each model is a certain distribution of means of the five dimensional Gaussian populations (eq. 2).

$$f[\vec{x}_j) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{\vec{x}_j - M_j}{\sigma}\right)^2} \quad (2)$$

where $\vec{x}_j$ is the input vector, $\vec{x} = [x_1, x_2, x_3, x_4, x_5]$.

In first model M_j is distributed in 0-5 interval according to exponential low, in second and third models the it is distributed by Gaussian and power lows respectively (eq. 3,4,5).

$$M_j(1) = \ln\left(\frac{1}{e^{-b2}} - RND\,M_j\left(e^{-b2} - e^{-b1}\right)\right); \qquad (3)$$

$$M_j(2) = N(\mu,\sigma); \qquad (4)$$

$$M_j(3) = RND\,M_j^{\frac{1}{\alpha+1}}(b2-b1)+b1; \qquad (5)$$

Where $\alpha = 2, b2 = 5, b1 = 0, \mu = 2.5, \sigma = 1, j = 1,10000$

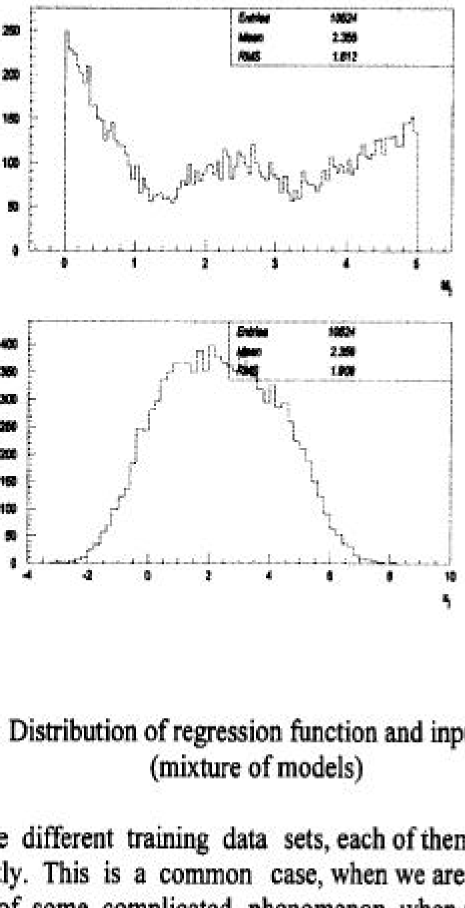

Figure 3: Distribution of regression function and input parameters (mixture of models)

So, we have three different training data sets, each of them represents the general data set only partly. This is a common case, when we are forced to use a Monte-Carlo simulation of some complicated phenomenon, when no a priory information can be assumed to describe exactly the underlying multidimensional probability distributions. Three different networks trained by these three models are applied to estimate the true regression function of the models mixture and the results are compared with the committee results.

The model to be considered is the mixture of these three models (control data set), in 0-1.7 – exponential distribution 1.7-3.3 – Gaussian distribution and 3.3-5 – power distribution (eq.6), (fig. 1).

$$M_j = \begin{cases} \ln\left(\frac{1}{e^{-b2}} - RN\,D\,M_j\left(e^{-b2} - e^{-b1}\right)\right) \; ; \; if \quad 0 \le M_j \\ \mathrm{N}(\mu,\sigma) ; \text{ if } \quad 1.7 < \mathrm{M_j} \le 3.3 \\ \mathrm{RN\,D\,M_j}^{\frac{1}{a+1}}(\mathrm{b2 - b1}) + \mathrm{b1} \; ; \text{ if } \quad 3.3 < \mathrm{M_j} \le 5 \end{cases} \tag{6}$$

On the figure 2 the estimation results by three different networks are plotted. It is easy to see that even for the best network there is significant bias of the estimation error. Comparing these results with ones in figure 3, where the estimation errors for different networks and committee are plotted, one can see, that the committee procedure leads to removing of the bias of estimation error and reduction of its variance. From table 1 one can see that the reduction of estimation error is about 10% and the median estimator gives better result than mean estimator.

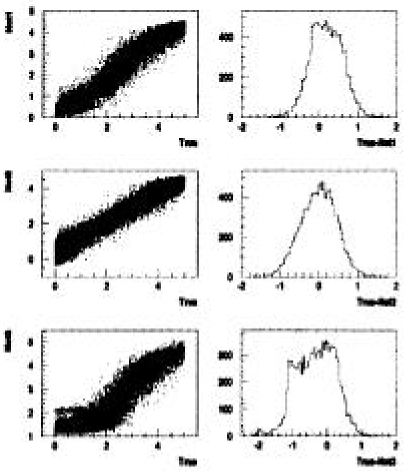

Figure 4: The regression function estimation by different networks trained on different models:
1 - exponential, 2 - Gaussian,
3 - power.

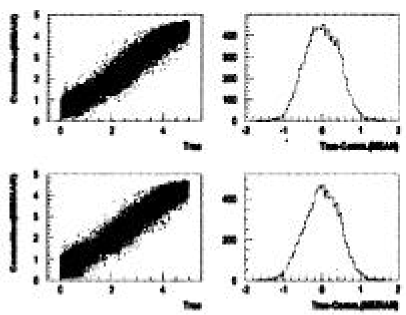

Figure 5: Mean and median committee results

Table 1: Root mean square errors of 3 different model networks and committee

	Net1	Net2	Net3	Median	Mean
RMSD	0.4536	0.4584	0.5688	0.4168	0.4455

Estimation of energy of primary cosmic ray flux Particle :

Such a selection of training and control data samples in toy problem is caused by a big importance of implementation of committee procedure for the Cosmic Ray Physics data analysis, where one has to solve an inverse problem and Monte Carlo simulation of propagation of cosmic ray radiation through the atmosphere and experimental installations is used. This is the simulation of development of extensive air showers (EAS) (Chilingaryan et al. 1997, Roth et al. 1997) in atmosphere induced by different primary nucleus. The problem to be solved can be presented as follows:

Simulation data *Experimeental data*

$$E < A(N_e, N_\mu, S, N_h, ...) \Rightarrow ?, ?(N_e, N_\mu, S, N_h, ...)$$

where the E and A are the primary energy and nuclei type inducing the extensive air shower in the atmosphere, N_e -number of electrons, N_μ -number of muons, ... are the secondary products (EAS components) of the cosmic radiation and atmosphere nucleus strong interactions. In the left side of above formula we prede_ne E and A and using Monte-Carlo method calculate the set of EAS characteristics, in the right side we have an experimentally measured EAS characteristics, but the primary type and energy are unknown. The problem is using the simulated data to reconstruct the primary nuclei energy and its type.

Our simulations are not exact representation of the real phenomena due to the many uncertainties and assumptions on the strong interaction parameters at energies non achievable for man-made accelerators.

The committee procedures were applied for the primary energy estimation of EAS simulated data, as we use \neural regression" for this problem (Chilingaryan A., H. Rebel 1997) In fig. 5 one can see the results of the primary energy estimation by different single networks and the committee of these networks. From this figure and table 2 it is easy to see, that the improvement of estimation error is significant, it is more than 10% comparing with the best single network among five used, and again median estimate is better than the mean.

Table 2: Root mean square errors and relative errors of 5 different networks and committee

	Net1	Net2	Net3	Net4	Net5	Median	Mean
RMSD	0.3802	0.4882	1.1378	0.5423	0.5999	0.3421	0.3586
Rel. Err.	0.0213	0.0262	0.0835	0.0286	0.0345	0.0197	0.0200

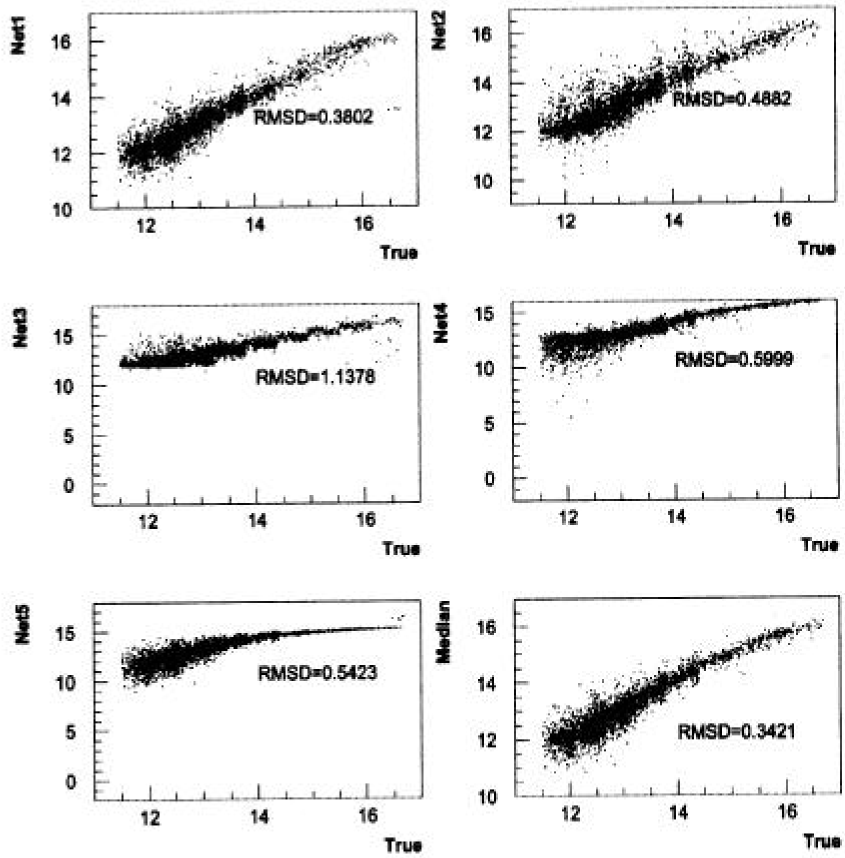

Figure 6: Estimation by different networks and their committee

Conclusions:

Using the committee of networks approach and the simulated data we demonstrate, that the reduction of estimation error is generally obtained even in case of small number of committee members. The improvement by about 10% can be viewed as a significant improvement, since the correlation between errors of different networks is rather high. If there are several trained networks to find the best one for a problem solution, it is always sensible to make a committee of that networks.

Refrences:

1. Bishop C.M., *Neural networks for pattern recognition*, Oxford University Press, New York (1995).
2. Chilingaryan, A., *Statistical Decisions Under Nonparametric A Priori Information*, Computer Physic Communications, 54 (1989), 381-390.
3. Chilingaryan, A., for KASCADE Collaboration, S.Ter-Antonyan, A.Vardanyan, *How to infer the mass composition from EAS observations demonstarated with KASCADE data*, 25 ICRC, OG 6.1.36, 105, Durban, (1997).
4. Roth, M., for KASCADE collaboration, S.Ter-Antonyan, A.Vardanyan, *How to infer the primary energy spectrum from EAS observations demonstarated with KASCADE data*, 25 ICRC, OG 6.2.13, Durban, (1997).
5. Chilingaryan A., H.Rebel, et al, *On the nonparametric classification and regression methods for the multivariate EAS data analysis*, Nuclear. Phys. B, 52B, (1997), 237.