Who firstly theoretically studied and proposed "Many could be better than all", Zhou et al or Perrone and Cooper?

David Ng

1 Introduction

Given a real valued function $f(x)$ and its $(N+1)$ approximates $f_1(x), \ldots, f_N(x), f_{N+1}(x)$, a question in the area of machine learning is: under what condition,

$$\hat{f}_{N+1} = \frac{\sum_{i=1}^{N+1} f_i}{N+1}$$

is worse (or better) than

$$\hat{f}_N = \frac{\sum_{i=1}^{N} f_i}{N}$$

in approximating f.

[1][2] addressed this question and claimed that their (theoretical) analysis "...reveals that that ensembling a selective subset of individual networks is superior to ensembling all the individual networks in some cases" (in Abstract of [2]), i.e., "many could be better than all" as the title of [2] says. [1] received the **Best Student Paper Award** in IJCAI2001. As Zhou admitted in a post (http://www.xys.org/xys/ebooks/others/science/dajia10 in www.xys.org, they were invited to submit an extended version of [1] to the journal of Artificial Intelligence and Journal of Computational Intelligence and Applications after IJCAI2001, and they accepted both invitations and published two similar papers [2] and [3]. [2] has been cited by more than 350 times since its publication. Many researchers believe that Zhou firstly studied this question and proposed 'many could be better than all". Zhou et al have also reiterated in their later publications that it was them who proposed "Many-all". For example, in [4] published this year, his co-worker and he stated that "Zhou et al. [24] **analyzed** the relationship between ensemble and its component learners from the context of both regression and classification, and **proved** that it may be better to combine many instead of all of the learners."

This note is to show that it is Perrone and Leon who firstly proposed and studied this question in [5] and explicitly claimed that averaging a subset of learners can be better than averaging all the learners, i.e. "many-all". This note also provides evidences that Zhou et al had read [5] before submitting [1].

2 Approach in [5]

To make the question at the outset of Introduction very-defined. [5] assumes that each approximation error $m_i = f - f_i$ is a random variable with zero mean, and denotes the correlation between m_i and m_j by $C_{ij} = E(m_i m_j)$ for $1 \leq i, j \leq N+1$.

[5] uses the mean square error (MSE) to measure the approximation quality of an approximate.

$$MSE(\hat{f}_N) = E[(\frac{1}{N}\sum_{i=1}^{N} m_i)^2] = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} E(m_i m_j) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} C_{ij}$$

Therefore, the question becomes: what is the necessary and sufficient condition for

$$MSE(\hat{f}_{N+1}) < MSE(\hat{f}_N)$$

Actually, the major contribution of [5] was to propose this question and formalize it in the probability language. Now it is very trivial to work out the condition.

[5] points out (the last inequality in Section 6) that the condition is:

$$(2N+1)MSE(\hat{f}_N) > 2\sum_{i\neq N+1} E[m_{N+1}m_i] + E(m_{N+1}^2) \tag{1}$$

Note that $N+1 = new$ in the original inequality in [5], we make this replacement to make this note easier to follow.

Although [5] does not call (1) "the necessary and sufficient condition". But [5] has said it to this effect in the paragraph just after this inequality: "...if a network does not satisfy this criterion ....".

Did [5] realize that "many is better than all" under some condition? The answer is yes. Actually, [5] clearly stated in the second last paragraph in Section 6 that "... adding more nets to the population is a waste of resource since it will not improve the performance....".

[5] was published in 1993.

3 Approach in [1][2]

[1][2] use the exact same way to make the above question very-defined. The only difference between [1][2] and [5] is some notations.

[1][2] said to the effect that

$$MSE(\hat{f}_N) \geq MSE(\hat{f}_{N-1})$$

under the following condition ((20) in [1], (17) in [2] , which was called "constraint" in [1][2]):

$$(2N-1)\sum_{i=1}^{N}\sum_{j=1}^{N} C_{ij} \leq 2N^2\sum_{i=1}^{N-1} C_{iN} + N^2 E[m_N^2] \tag{2}$$

The only difference between the above inequality and (17) in [2] (i.e., (20) in [1]) is that we label the new "ensemble" by N instead of k to make the inequality more readable.

It is trivial that

$$\sum_{i=1}^{N}\sum_{j=1}^{N} C_{ij} = \sum_{i=1}^{N-1}\sum_{j=1}^{N-1} C_{ij} + 2\sum_{i=1}^{N-1} E[m_i m_N] + E[m_N^2]$$

$$\sum_{i=1}^{N-1}\sum_{j=1}^{N-1} C_{ij} = (N-1)^2 MSE(\hat{f}_{N-1})$$

Therefore (2) is equivalent to

$$(2N-1)\{(N-1)^2 MSE(\hat{f}_{N-1})+2\sum_{i=1}^{N-1} E[m_i m_N]+E[m_N^2]\} \leq 2N^2 \sum_{i=1}^{N-1} C_{iN}+N^2 E[m_N^2]$$

Then

$$(2N-1)MSE(\hat{f}_{N-1}) \leq 2\sum_{i=1}^{N-1} E[m_i m_N] + E[m_N^2] \quad (3)$$

Comparing (2)(3) and (1), it is obvious that (3) (i.e. (2)) is just the "$\Leftarrow$" version of (1). Actually, (2) must be so, otherwise, it would be wrong.

Zhou claimed in [2] that "we get" (2) and "reach the conclusion that many ... may be .. better than... all" based on (2).

[1] did not cite [5], and [2] cited [5] but did not acknowledge [5] in the part on (2).

Did Zhou et al not read [5] carefully and independently find (2)? The answer is no, since Zhou et al published a Chinese version of [1] and acknowledged [5] on (2) before IJCAI2001 in [6].

4 Conclusions

The answer to the question set in the title is " Perrone and Cooper". it is also evident that the theoretical result of Zhou et al on "many-all" were directly taken from [5]. This academic misconduct was carried out intentionally.

Acknowledgement

Several anonymous authors in www.xys.org found the relationship between (1) and (2) and accused Zhou et al of their academic misconduct on "many-all" before me.

References

[1] Z-H Zhou, J-X Wu, Y. Jiang and S-F Chen, Genetic Algorithm based Selective Neural Network Ensemble, IJCAI01, pp. 797-802.

[2] Z-H Zhou, J Wu, and W. Tang. Ensembling Neural Networks: Many could be Better than All, AIJ 2002, 137(1-2):239-263.

[3] Z-H Zhou, J-X Wu, W. Tang, Z-Q Chen, Combing regression estimators: GA-based selectively neural network ensemble. International Journal of Computational Intelligence and Applications, 2001, 1(4): 341-356.

[4] N. Li, Z-H Zhou. Selective Ensemble Under Regularization Framework. MCS'09, LNCS 5519, 293-303.

[5] M. P. Perrone, L. N. Cooper. When Networks disagree: ensemble method for neural networks. Artificial Neural Networks for Speech and Vision, 1993, 126-142.

[6] J-X. Wu, Z-H Zhou, et al, A Selective constructing approach to neural network ensemble, JOURNAL OF COMPUTER RESEARCH AND DEVELOPMENT (Chinese), 1039-1044, No.9, 2000.