

Ensemble of GA based Selective Neural Network

Ensembles

Jian-Xin WU Zhi-Hua ZHOU Zhao-Qian CHEN

National Laboratory for Novel Software Technology
Nanjing University
Nanjing, 210093, P.R.China
wujx@ai.nju.edu.cn {zhouzh, chenzq}@nju.edu.cn

Abstract

Neural network ensemble is a learning paradigm where several neural networks are jointly used to solve a problem. In this paper, e-GASEN, a two-layer neural network ensemble architecture is proposed, in which the base learners of the final ensemble are also ensembles. Experimental results show that e-GASEN generalizes better than a popular ensemble method. The reason why e-GASEN works is also discussed. We believe that the different layers of e-GASEN attain good generalization ability for different reasons. The first layer ensembles profit from the selected individual neural networks that are moderately divergent but generalize well, while the second layer ensemble profits from the divergency among the first layer ensembles.

1 Introduction

Since neural computing has no rigorous theoretical framework until now, whether a neural network based application will be successful or not is almost fully determined by the practitioner. In general, the more experienced the practitioner is, the more chances the application will have of being success. However, users are often with little knowledge on neural computing. Therefore the rewards that neural network techniques may return do not always appear.

In the beginning of the 1990's, Hansen and Salamon showed that the generalization ability of a neural network system can be significantly improved through ensembling individual neural networks, i.e. training several neural networks and combining their results in some way[1]. Later, Sollich and Krogh defined neural network ensemble as a collection of a (finite) number of neural networks that are trained for

the same task[2]. Since it behaves remarkably well and is very easy to use, neural network ensemble is regarded as a promising methodology that can benefit not only experts in neural computing but also ordinary engineers. And neural network ensemble has already been used in many real domains such as handwritten digit recognition[3], scientific image analysis[4], face recognition[5][6], OCR[7], seismic signal classification[8], etc.

Many works have been done to investigate why and how neural network ensemble works. The classical one is Krogh and Vedelsby[9]'s work, in which they derived the famous equation $E = \bar{E} - \bar{A}$. It clearly demonstrates that the generalization ability of the ensemble is determined by the average generalization ability and the average ambiguity (divergency) of the individual neural networks that constitute the ensemble.

Many ensemble methods have been proposed in the literature. The most attractive methods mainly include *simple ensemble*[2], *AdaBoost*[10], and *bagging*[11]. These methods combine outputs of all the base learners at hand. Usually, the base learner is a neural network or a classification tree.

If a base learner is of high generalization error and low ambiguity, adding it into the ensemble will definitely deteriorate the ensemble's generalization ability. However, there is no guarantee that such "bad" base learner will never appear. This means that in some circumstances using all the base learners at hand may not be the best choice.

GASEN (Genetic Algorithm based Selective ENsemble) was proposed in [12], which trains several neural networks and then employs genetic algorithm to select an optimum subset of those networks to constitute an ensemble. Experiments

show that GASEN is superior to *simple ensemble*, even if it tends to select only a small quantity of neural networks.

In this paper, we argue that if we employ a two-layer ensemble architecture, i.e. when the base learner itself is also an ensemble, the final ensemble will have better generalization ability. The architecture we used in this paper is a *simple ensemble* of GASEN. The final ensemble is composed of several GASENs, and the final ensemble's output is the average of all individual GASEN's outputs. This architecture is abbreviated as *e-GASEN*. By analyzing the experimental results, we believe that GASEN and *e-GASEN* promotes the final ensemble's generalization ability in different ways.

The rest of this paper is organized as follows. In Section 2, the equation $E = \bar{E} - \bar{A}$, i.e. relation between the generalization ability of the ensemble, the generalization ability of individual base learners and the average ambiguity of the base learners, is first explained. Then GASEN is briefly introduced. In Section 3, *e-GASEN* is proposed. Some experiments are also reported. In Section 4, The facts revealed by experiments are discussed. The reason about how and why such a two-layer ensemble architecture works is analysed. Finally in Section 5, conclusions are drawn and several issues for future work are indicated.

2 GASEN

Suppose the learning task is to use an ensemble that comprises N base learners to approximate a function $f: \mathbf{R}^m \rightarrow \mathbf{R}^n$. The predictions of the base learners are combined through *weighted averaging*, where a weight w_i ($i = 1, 2, \dots, N$) is assigned to the individual base learner f_i , and w_i satisfies equation (1) and (2):

$$0 < w_i < 1 \quad (1)$$

$$\sum_{i=1}^N w_i = 1 \quad (2)$$

The output of the ensemble is computed according to equation (3), where f_i is the output of the i -th base learner.

$$\bar{f}(x) = \sum_{i=1}^N w_i f_i(x) \quad (3)$$

For convenience of discussion, here we assume that each base learner has only one output component, i.e. the function to be approximated is $f: \mathbf{R}^m \rightarrow \mathbf{R}$. But note that it can be easily generalized to situations where each base learner has multiple output components.

Suppose $x \in \mathbf{R}^m$ is randomly sampled according to a distribution $p(x)$. The expected output for x is $d(x)$. Then the error $E_i(x)$ of the i -th base learner on input x and the error $E(x)$ of the ensemble on input x are respectively:

$$E_i(x) = (f_i(x) - d(x))^2 \quad (4)$$

$$E(x) = (\bar{f}(x) - d(x))^2 \quad (5)$$

Then the generalization error E_i of the i -th base learner on the distribution $p(x)$ and the generalization error E of the ensemble on the distribution $p(x)$ are respectively:

$$E_i = \int dx p(x) E_i(x) \quad (6)$$

$$E = \int dx p(x) E(x) \quad (7)$$

The average error of the base learners on input x is:

$$\bar{E}(x) = \sum_{i=1}^N w_i E_i(x) \quad (8)$$

Then the average generalization error of the base learners on the distribution $p(x)$ is:

$$\bar{E} = \int dx p(x) \bar{E}(x) \quad (9)$$

Accordingly, the ambiguity of the i -th base learner on input x , the ambiguity of the i -th base learner on the distribution $p(x)$, the average ambiguity of base learners on input x , and the average ambiguity of the base learners on the distribution $p(x)$ are defined respectively as:

$$A_i(x) = (f_i(x) - \bar{f}(x))^2 \quad (10)$$

$$A_i = \int dx p(x) A_i(x) \quad (11)$$

$$\bar{A}(x) = \sum_{i=1}^N w_i A_i(x) \quad (12)$$

$$\bar{A} = \int dx p(x) \bar{A}(x) \quad (13)$$

After a few algebraic manipulations, Krogh and Vedelsby reached the famous formula (14) which states that the generalization ability of the ensemble

is determined by the average generalization ability and the average ambiguity of the base learners that constitutes the ensemble.

$$E = \bar{E} - \bar{A} \quad (14)$$

Now we define the correlation between the i -th and the j -th individual base learner as:

$$C_{ij} = \int dx p(x) (f_i(x) - d(x))(f_j(x) - d(x)) \quad (15)$$

Then, according to [12] and [13], we have

$$E = \sum_{i=1}^N \sum_{j=1}^N w_i w_j C_{ij} \quad (16)$$

When the base learners are combined using the *simple ensemble* method, i.e. $w_i = 1/N$ for every i , we have

$$E = \sum_{i=1}^N \sum_{j=1}^N C_{ij} / N^2 \quad (17)$$

It is proved that when using the *simple ensemble* method and when formula (18) is satisfied, then omitting the k -th base learner will improve the ensemble's generalization ability [12].

$$(2N-1) \sum_{\substack{i=1 \\ i \neq k}}^N \sum_{\substack{j=1 \\ j \neq k}}^N C_{ij} < 2(N-1)^2 \sum_{\substack{i=1 \\ i \neq k}}^N C_{ik} + (N-1)^2 E_k \quad (18)$$

Now a conclusion is arrived that after the neural networks are trained, in some cases ensembling an appropriate subset of the neural networks is superior to ensembling all of them. The individual neural network that should be omitted satisfy equation (18).

This statement is also partly approved by Liu, Yao and Higuchi[14]. After trained several neural networks with *negative correlation learning* and evolutionary computation, they used the k -means algorithms to divide the individuals into different clusters. In every cluster, the fittest individual network is selected as a representative of the cluster. They compared the ensemble formed of these representatives and the ensemble formed of all the networks. No statistically significant difference is observed between them in their experiments. This observation implies that the ensemble does not have to use all the networks to achieve good performance.

GASEN is proposed based on this conclusion, which first trains several individual neural networks

independently and then employs genetic algorithm to select an optimum subset of individual networks to constitute an ensemble. The selected neural networks are combined together using simple averaging. However, experimental data show that GASEN is superior to using all the available networks at hand [12].

Although genetic algorithm is used in both methods of [14] and [12], they are quite different. In [14] genetic algorithm is used to evolve a population of neural networks that are negatively correlated, while in [12] genetic algorithm is used to select a subset of neural networks to constitute the ensemble.

3 e -GASEN

It is well known that in order for an ensemble to work well, the individual neural networks should respond as *independent* as possible to an input. If the independency requirement is satisfied, the ensemble's generalization error will decrease when more neural networks are added into it. However, the marginal error reduced by every newly added neural network tends to decrease when the ensemble grows larger and larger[13][15].

The GASEN method underwent a genetic algorithm based selection process. After selection, GASEN's size, i.e. the number of neural networks survived the selection process, is rather small. Experimental data in [12] show that if N neural networks are trained, averagely GASEN will select only about $N/4$ among them to form an ensemble.

The benefits brought by the genetical selection process is enjoyable. However, we believe that if more neural networks are included, in some cases the generalization error of the ensemble may be further reduced.

This is the motivation of e -GASEN, which is a natural extension of GASEN. Given a learning task, we may train several ensembles using the GASEN algorithm first. Then, an e -GASEN is formed by combining these GASENs by using the *simple ensemble* method, i.e. averaging the output of several GASENs' on an input to form the e -GASEN's output.

The e -GASEN is a two-layer ensemble architecture. Since e -GASEN is formed by averaging several GASENs and every GASEN is constructed by averaging several single neural networks, e -GASEN

Table 1 Experimental results on *simple ensemble*, GASEN, and *e*-GASEN

Data set	<i>simple ensemble</i>		GASEN		<i>e</i> -GASEN	
	error	deviation	Error	deviation	error	deviation
<i>Friedman#1</i>	1.33	0.35	0.49	0.15	0.36	0.048
<i>Boston Housing</i>	12.25	1.06	10.71	0.62	10.10	0.33
<i>Ozone</i>	22.85	1.30	20.42	1.66	19.18	0.47
<i>Servo</i>	0.22	0.035	0.25	0.08	0.21	0.036

Table 2 The mean error-ambiguity decomposition of generalization error

Data set	<i>simple ensemble</i>			GASEN			<i>e</i> -GASEN		
	E	\bar{E}	\bar{A}	E	\bar{E}	\bar{A}	E	\bar{E}	\bar{A}
<i>Friedman#1</i>	1.33	2.97	1.64	0.49	1.10	0.61	0.36	1.14	0.79
<i>Boston Housing</i>	12.25	18.11	5.86	10.71	14.37	3.66	10.10	14.09	3.99
<i>Ozone</i>	22.85	32.75	9.90	20.42	26.05	5.63	19.18	25.94	6.76
<i>Servo</i>	0.22	0.48	0.25	0.25	0.40	0.14	0.21	0.39	0.18

may be viewed as averaging some *selected* single neural networks. So we may define the size of an *e*-GASEN as the number of single neural networks contained in it. In this sense, the size of an *e*-GASEN equals the sum of sizes of all its component GASENs.

We use four regression problems that were used in [12] to compare the performance of *simple ensemble*, GASEN and *e*-GASEN.

The first problem is *Friedman#1* proposed by Friedman [16]. There are 5 continuous attributes. The data set is generated according to equation (19) where the noise item ε satisfies normal distribution $N(0, 1)$ and x_i ($i = 1, 2, \dots, 5$) satisfies uniform distribution $U[0, 1]$. In our experiments the size of the training set and the test set are respectively 200 and 1000.

$$t = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \varepsilon \quad (19)$$

The second problem is *Boston Housing* from UCI machine learning repository[17]. There are 11 continuous attributes and 1 categorical attribute. The data set comprises 506 examples among which 400 examples make up the training set and the rest 106 examples make up the test set in our experiments.

The third problem is *Ozone* proposed by Breiman and Friedman [18]. There are 9 continuous attributes. The data set comprises 366 examples. Since the intention of the experiments is not to compare the ability of dealing with missing values, 1 attribute and 36 examples that has missing values are omitted.

Therefore in our experiments there are 8 continuous attributes and 330 examples among which 250 examples make up the training set and the rest 80 examples make up the test set.

The fourth problem is *Servo* from UCI machine learning repository. There are 4 categorical attributes. The data set comprises 167 examples among which 130 examples make up the training set and the rest 37 examples make up the test set in our experiments. Note that some researchers [19] believe that this problem is very difficult because it involves some kind of extreme nonlinearity.

For each problem we use *bagging* on the training set to generate 20 single-hidden-layered BP networks. The *simple ensemble* method is formed by averaging these networks. After performing genetical selection, a GASEN is constructed by averaging the selected networks. For every problem we perform 20 runs and record the average mean squared error and the standard deviations of these errors on the test set. An *e*-GASEN is formed by averaging 4 GASENs. So there are totally 5 runs of *e*-GASEN. The average mean squared error and corresponding standard deviation is also recorded. Experimental results are shown in Table 1.

Statistical tests show that on the *Friedman#1*, *Boston Housing*, and *Ozone* data sets, GASEN's generalization error is significantly lower than that of the *simple ensemble* method, and *e*-GASEN attains still lower generalization errors than GASEN. On the

Servo data set, GASEN is slightly inferior to *simple ensemble*. The *e*-GASEN method's performance, however, has no significant difference with that of the *simple ensemble* method.

From the aforementioned statistics we may conclude that *e*-GASEN is superior to both GASEN and *simple ensemble*.

4 Discussion

Until now, we are not very clear through what mechanism *e*-GASEN works so well.

Following formula (14), the generalization error of an ensemble (E) can be decomposed as difference of the mean error part \bar{E} and the mean ambiguity part \bar{A} . The mean error-ambiguity decomposition of *simple ensemble*, GASEN and *e*-GASEN on the four data sets are tabulated in Table 2. The mean error part \bar{E} is calculated by averaging the error of the individual neural networks that consist the ensemble on the test set. Then, The mean ambiguity part \bar{A} is calculated from formula (14) with the help of \bar{E} .

It is clear that the mean error part \bar{E} of GASEN is quite smaller than that of *simple ensemble*. The mean ambiguity part \bar{A} of GASEN is also smaller than that of *simple ensemble*, the decrease in \bar{E} is more significant. It means that GASEN may attain better generalization ability by selects base learners that are of only *moderate* ambiguity but are well-generalized. This analysis is somewhat different from our previous one in [12], in which we believe that GASEN's genetical selection process will increase ambiguity.

Since every GASEN has a relatively small generalization error, *e*-GASEN is hard to get smaller error by further lowering the mean error part \bar{E} . From Table 2 we can find that the mean error of GASEN and *e*-GASEN has no obvious difference, while *e*-GASEN has a higher mean ambiguity. Therefore we believe that *e*-GASEN mainly profits from the divergence among different GASENs.

5 Conclusions and future work

In this paper, we re-examined the GASEN, i.e. Genetic Algorithm based Selective ENsemble method and proposed *e*-GASEN, a natural extension of GASEN, which combines several GASENs by using

simple averaging. Through analyses of experimental results, a conjecture on how and why *e*-GASEN works is proposed. We believe that GASEN works by selecting neural networks that are of only *moderate* ambiguity but are well-generalized. And, the *e*-GASEN method gains from the divergence among different GASENs. However, analyses presented in this paper are very preliminary. More experiments and theoretical works are still needed to clarify the rules behind GASEN and *e*-GASEN. For theoretical analysis, we believe that the bias-variance decomposition may be helpful [20]. Moreover, whether other kinds of ensembles, e.g. weighted averaged one, can play the roles of GASEN in *e*-GASEN, is an interesting issue for future exploration.

Acknowledgements

The National Natural Science Foundation of P. R.China and the Natural Science Foundation of Jiangsu Province, P. R.China, supported this research.

References

- [1] L. K. Hansen and P. Salamon, "Neural network ensembles," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 12, no. 10, pp. 993-1001, 1990.
- [2] P. Sollich, A. Krogh, "Learning with Ensembles: How over-fitting can be useful," In Advances in Neural Information Processing Systems 8, pp. 190-196, 1996.
- [3] L. K. Hansen, L. Liisberg, and P. Salamon, "Ensemble methods for handwritten digit recognition," In Proc. IEEE-SP Workshop on Neural Networks for Signal Processing, pp. 333-342, 1992, IEEE Computer Society.
- [4] K. J. Cherkauer, "Human expert level performance on a scientific image analysis task by a system using combined artificial neural networks," In Proc. 13th AAAI Workshop on Integrating Multiple Learned Models for Improving and Scaling Machine Learning Algorithms, pp. 15-21, 1996, AAAI.
- [5] S. Gutta and H. Wechsler, "Face recognition using

- hybrid classifier systems,” In Proc. IEEE Int. Conf. on Neural Networks, pp. 1017-1022, 1996, IEEE Computer Society.
- [6] F. J. Huang, Z.-H. Zhou, H.-J. Zhang, and T. Chen, “Pose invariant face recognition,” In Proc. 4th IEEE Int. Conf. on Automatic Face and Gesture Recognition, pp. 245-250, Grenoble, France, 2000, IEEE Computer Society.
- [7] J. Mao, “A case study on bagging, boosting and basic ensembles of neural networks for OCR,” In Proc. IEEE Int. Joint Conf. on Neural Networks, vol.3, pp. 1828-1833, 1998, IEEE Computer Society.
- [8] Y. Shimshoni and N. Intrator, “Classification of seismic signals by integrating ensembles of neural networks,” IEEE Trans. Signal Processing, vol. 46, no. 5, pp. 1194-1201, 1998.
- [9] A. Krogh and J. Vedelsby, “Neural network ensembles, cross validation, and active learning,” In Advances in Neural Information Processing Systems 7, pp. 231-238, 1995, MIT.
- [10] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119-139, 1997.
- [11] L. Breiman, “Bagging predictors,” Machine Learning, vol. 24, no. 2, pp. 123-140, 1996.
- [12] Z.-H. Zhou, J.-X. WU, Y. Jiang, and S.-F. Chen, “Genetic Algorithm based Selective Neural Network Ensemble,” to appear in Proc. IJCAI’01, Seattle, USA, 2001.
- [13] M. P. Perrone, L. N. Cooper, “When networks disagree: Ensemble method for neural networks,” In Mammone R.J. eds. Artificial Neural Networks for Speech and Vision, London, Chapman-Hall, pp. 126-142, 1993.
- [14] Y. Liu, X. Yao and T. Higuchi, “Evolutionary Ensembles with Negative Correlation Learning,” IEEE Trans. Evolutionary Computation, vol. 4, no. 4, pp. 380-387, 2000.
- [15] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, “Boosting the margin: A new explanation for the effectiveness of voting methods,” The Annals of Statistics, vol. 26, no. 5, pp. 1651-1686, 1998.
- [16] J. Friedman, “Multivariate adaptive regression splines (with discussion),” Annals of Statistics, vol. 19, no. 1, pp. 1-141, 1991.
- [17] C. Blake, E. Keogh, and C. J. Merz, UCI Repository of machine learning databases [<http://www.ics.uci.edu/~mllearn/MLRepository.html>], Dept. of Information and Computer Science, University of California, Irvine, California, 1998.
- [18] L. Breiman and J. Friedman, “Estimating optimal transformations in multiple regression and correlation (with discussion),” Journal of the American Statistical Association, vol. 80, pp. 580-619, 1985.
- [19] J. R. Quinlan, “C4.5: Programs for Machine Learning,” Morgan Kaufmann, San Mateo, California, 1993.
- [20] L. Breiman, “Bias, variance and arcing classifiers,” Technical Report 460, Statistics Department, University of California at Berkeley, 1996. (<ftp://ftp.stat.berkeley.edu/users/breiman/arcall.ps.Z>)