Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

Clustering-based selective neural network ensemble

FU Qiang (傅 强)[†1], HU Shang-xu (胡上序)[1], ZHAO Sheng-ying (赵胜颖)[2]

([1]*Laboratory of Intelligence Information Engineering, Zhejiang University, Hangzhou 310027, China*)

([2]*UTStarcom Telecom Ltd., Hangzhou 310027, China*)

[†]E-mail: fuqiang@zju.edu.cn

Received Mar. 27, 2004; revision accepted Dec. 1, 2004

Abstract: An effective ensemble should consist of a set of networks that are both accurate and diverse. We propose a novel clustering-based selective algorithm for constructing neural network ensemble, where clustering technology is used to classify trained networks according to similarity and optimally select the most accurate individual network from each cluster to make up the ensemble. Empirical studies on regression of four typical datasets showed that this approach yields significantly smaller ensemble achieving better performance than other traditional ones such as Bagging and Boosting. The bias variance decomposition of the predictive error shows that the success of the proposed approach may lie in its properly tuning the bias/variance trade-off to reduce the prediction error (the sum of bias^2 and variance).

Key words: Neural network, Ensemble, Clustering
doi:10.1631/jzus.2005.A0387 **Document code:** A **CLC number:** TP387

INTRODUCTION

Neural network ensemble is becoming a hot spot in machine learning and data mining recently. Many researchers have shown that simply combining the output of many neural networks can generate more accurate predictions than that of any of the individual networks. Most previous work either focused on how to combine the output of multiple trained networks or how to directly design a good set of neural networks.

Theoretical and empirical work showed that a good ensemble is one where the individual networks have both accuracy and diversity, namely the individual networks make their errors on difference parts of the input space (Hansen and Salamon, 1990; Krogh and Vedelsdy, 1995). Many approaches have been proposed to construct such ensembles. One group of these methods obtains diverse individuals by training accruate networks on different training set, such as bagging, boosting, cross validation and using artificial training examples (Breiman, 1996; Schapire, 1990; Krogh and Vedelsdy, 1995; Melville and Mooney, 2003). Another group of these methods adopts different topologies, initial weigh setting, parameter setting and training algorithm to obtain individuals. For example, Rosen (1996) adjusted the training algorithm of the network by introducing a penalty term to encourage individual networks to be decorrelated; Liu and Yao (2000) used negative correlation learning to generate negatively correlated individual neural network. The third group is named selective approach group where the diverse components are selected from a number of trained accurate networks. For example, Opitz and Shavlik (1996) proposed a generic algorithm to search for a highly diverse set of accurate networks; Lazarevic and Obradoric (2001) proposed a pruning algorithm to eliminate redundant classifiers; Navone *et al.*(2000) proposed another selective algorithm based on bias/variance decomposition; GASEN proposed by Zhou *et al.*(2001) and PSO based approach proposed by Fu *et al.*(2004) also were introduced to select the ensemble components.

This paper proposes a new selective algorithm based clustering technology. After a number of neural networks are trained, *k*-means clustering is used to divide them into some clusters based on the output of all networks on the same input. One most accurate network in each cluster is selected to join the ensemble. The proposed method applied to several datasets indicated that this approach yields significantly smaller size ensemble achieving much better performance.

METHOD

In order to improve the prediction accuracy achieved by an ensemble, we need to ensure accuracy of networks and diversity between individuals. The accuracy can be described by the mean square error and achieved by proper training algorithms of neural network. However, although discussed by many researches, the diversity of ensemble has no standard definition so far. The “diversity” assumption means that the networks have to make independent prediction errors. So, considering the output of networks on the same input dataset, we can commonly agree that the more different the output between the individuals is, the more diverse the ensemble is.

The diversity can be achieved by selecting some members from many accurately trained networks. Given a set $\boldsymbol{H}$ of all trained networks h_t, t=1,…,T, our goal is to select some of them to make up the ensemble. Considering the difficulty of selecting diversity and accuracy at the same time, we can apply an easier method to gradually achieve the diversity and accuracy. First, we employ clustering technology to divide all networks into some groups (clusters) according to similarity of the networks. Then, one most accurate individual in each group on the validation set is selected. Finally, all selected individuals construct the ensemble.

Let $h_i(x)$ be the prediction that the ith network makes for the instance $x\in\boldsymbol{S}$. It is apparent that what the network h_i makes for the entire training set $\boldsymbol{S}$ can be represented as a vector $\boldsymbol{Y}_i$ containing m prediction values, one for each of data example from the training set $\boldsymbol{S}$.

Consider all T prediction vectors $\boldsymbol{Y}_t$ that T networks make. Each of these vectors $\boldsymbol{Y}_t$ may be treated as a data pattern with m attributes. Therefore a clustering algorithm can be applied to the set that contains T patterns, with m attributes each.

Standard k-means algorithm is employed to cluster the set of neural networks represented by the data pattern $\boldsymbol{Y}$. Our goal is to divide data pattern $\boldsymbol{Y}$ $=\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T\}$ into k clusters $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_k$, where the size of cluster $\boldsymbol{D}_i$ is n_i and the mean of the data in cluster $\boldsymbol{D}_i$ is $\boldsymbol{\mu}_i$, the distance between two vectors is defined by the Euclidian distance. So clustering can be achieved by finding μ_i which make

$$J_e=\sum_{i=1}^{k}\sum_{y\in\boldsymbol{D}_i}\|y-\mu_i\|^2 \qquad (1)$$

minimized. The points $\boldsymbol{\mu}_i$ are known as cluster centroids or cluster means.

The standard k-means algorithm is shown in Table 1.

Table 1 Standard k-means algorithms

```
Input: number of clusters $k$
Procedure:
    1. Initialize $k$ means vectors at random, $\boldsymbol{\mu}_i$, ($i$=1, 2, …, $k$)
    2. Classify the input vectors according to the closest
       means vectors $\boldsymbol{\mu}_i$, to $k$ clusters $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_k$
    3. Re-compute $\boldsymbol{\mu}_i$,
    4. If there are any changes in each $\boldsymbol{\mu}_i$, for all input vectors,
       return to Step 2. Otherwise stop.
Output: $k$ cluster centroids $\boldsymbol{\mu}_i$, and $k$ clusters $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_k$
```

Obviously after clustering the diversity between networks in different groups is greater than those in the same group. We can maintain the diversity by choosing the most accurate networks in each group to make up the ensemble.

In k-means algorithm, cluster number k must be determined in advance. To confirm the best k value, we can compare the ensemble prediction error on validation set and choose the best ensemble to determine the corresponding k value.

The clustering-based selective neural network ensemble algorithm is shown in Table 2 (see the next page).

EXPERIMENTAL RESULTS

We use four regression problems to compare the performance of clustering-based approach and two main ensemble approaches, i.e. Bagging and Boosting.

The first problem is Friedman #1 proposed by Friedman *et al.*(1983). There are 5 continuous attribu-

Table 2 Algorithm of clustering based-selective neural network ensemble

```
Input: training set $\boldsymbol{S}$, validation set $\boldsymbol{V}$, trained neural networks
    $h_t\,(t=1, 2, \ldots, n)$,
Procedure:
    for $i$=1 to $n$ {
    $\boldsymbol{Y}_i=h_i\,(\boldsymbol{S})$
    }
    for $j$=1 to $k$ {
    Create group $\boldsymbol{D}_j$ by clustering $\boldsymbol{Y}_i$, $i$=1, 2, ..., $n$, to it
    }
    for $j$=1 to $k$ {
    Compute accuracy of each network in $\boldsymbol{D}_j$ on the validation set
    Select the most accurate network in $\boldsymbol{D}_j$ and join the ensemble $\boldsymbol{E}^*$.
    }
Output: ensemble $\boldsymbol{E}^*$
    $\boldsymbol{E}^*(S) = \text{ave}\sum_{h\in\boldsymbol{E}^*} h(S)$
```

tes. The dataset is generated according to Eq.(2) where x_i (i=1, 2, …, 5) satisfies uniform distribution $U(0,1)$ and the noise item satisfies normal distribution $N(0,1)$. The size of the dataset in our experiments is 2800.

$$y=10\sin(\pi x_1x_2)+20(x_3-0.5)^2+10x_4+5x_5+N(0,1)\quad x_i\sim U(0,1) \tag{2}$$

The second problem is Plane proposed by Ridgeway *et al.*(1999). There are 2 continuous attributes. The dataset is generated according to Eq.(3) where x_i (i=1, 2) satisfies uniform distribution $U(0,1)$ and the noise item satisfies normal distribution $N(0,0.05)$. The size of the dataset in our experiments is 1000.

$$y=0.6x_1+0.3x_2N(0,0.05)\quad x_i\sim U(0,1) \tag{3}$$

The third problem is Boston Housing from UCI machine learning repository (Blake *et al.*, 1998). There are 11 continuous attributes and 1 categorical attribute. The dataset is comprised of 506 examples.

The fourth problem is Ozone proposed by Breiman and Friedman (1985). There are 9 continuous attributes. The dataset is comprised of 366 examples. We use 330 examples and 8 attributes after omitting 1 attribute and 36 examples with missing values.

In our experiments, 10-fold cross validation is employed on each dataset to compare the prediction error of the clustering-based approach and bagging. The result of the approach is the average result of ten folds. The dataset is divided into 10 subsets among which one subset is used as test set and the other 9 subsets make up the training set in each fold in turn. In each fold, the training set is bootstrap sampled from the training set of the fold. The size of the training set is about 80% of that of the fold, and the remaining 20% is used as validation set. In each fold, 20 neural networks are trained and clustering-based approach is employed to select some individual members to construct the ensemble.

The neural networks in the ensembles are trained by implementation of the back propagation algorithm in MATLAB. Each network has one hidden layer that is comprised of 10 units. The parameters such as the learning rate are set to default values of MATLAB.

As comparisons, Bagging algorithm (Breiman, 1996) and Boosting algorithm are respectively performed on each dataset in the same condition. We employ Adaboost.R2 (Drucker, 1999) to deal with regression problems. Obviously, for Bagging and Boosting each ensemble contains 20 neural networks and for the Clustering based approach the number of networks is far less than twenty.

Table 3 shows that in most problems (Friedman #1, Boston Housing, Ozone), the clustering based approach is significantly better than both Bagging and Boosting algorithm; no improvement is shown in Plane partly because of the simplicity of the problem; in all problems, the number of networks in the ensem-

Table 3 Comparison of the prediction error of methods on regression

Dataset	Bagging		Boosting		Clustering based approach	
	Prediction error	Number of networks	Prediction error	Number of networks	Prediction error	Number of networks
Friedman #1	1.172	20	1.271	20	1.162	6
Plane	0.0027	20	0.0027	20	0.0027	5.1
Boston Housing	42.60	20	36.16	20	32.47	6.3
Ozone	21.83	20	21.97	20	18.81	4.3

ble made by the proposed approach is far less than that done by Bagging and Boosting (about 75% reduction). In Friedman #1 problem, Bagging is better than Boosting; and in Boston Housing problem, Boosting is better than Bagging. In other problems (Ozone, Plane), Boosting is similar to Bagging.

We also report the performance of the clustering based approach when the number of clusters *k* changes. The number of networks in the ensemble is equal to that of clustering because one best network in each cluster is selected to join the ensemble in our algorithm. Fig.1 shows that no matter how many networks (*k*>1) are in an ensemble, the performance is always better than even the best single network (*k*=1); Fig.1 also shows that the best *k* value to minimize prediction error is far less than the number of all networks (20). It means that the best ensemble need not employ all neural networks.

DISCUSSION

To find the reason for the success of the proposed approach, bias-variance decomposition is employed to analyze the results of Bagging, Boosting and clustering based approach. Bias-variance decomposition (German *et al.*, 1992; Hansen, 2000) is a powerful tool from sampling theory for analyzing the working mechanism of supervised learning approaches. Bias-variance decomposition for regression (quadratic loss) avers that the prediction error of an estimator can be broken down into two components: bias^2 (or bias) and variance. The bias measures show how closely the average estimate of the learning approach matches the target. The variance measures show how much the estimate of the learning approach fluctuates the different training sets of the given size. These two usually work in opposition to each other: attempts to

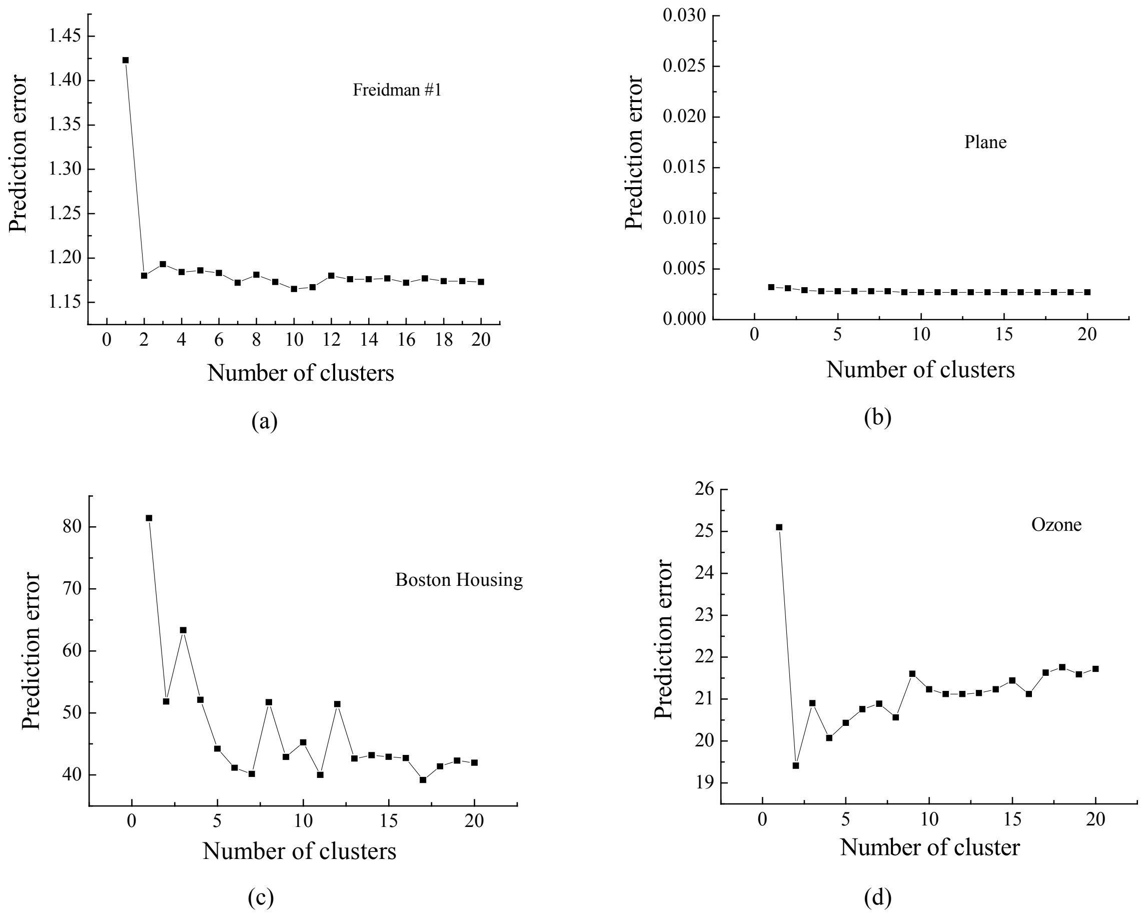


Fig.1 Comparison of prediction error on different number of clusters
(a) Friedman #1 problem; (b) Plane problem; (c) Boston Housing problem; (d) Ozone problem

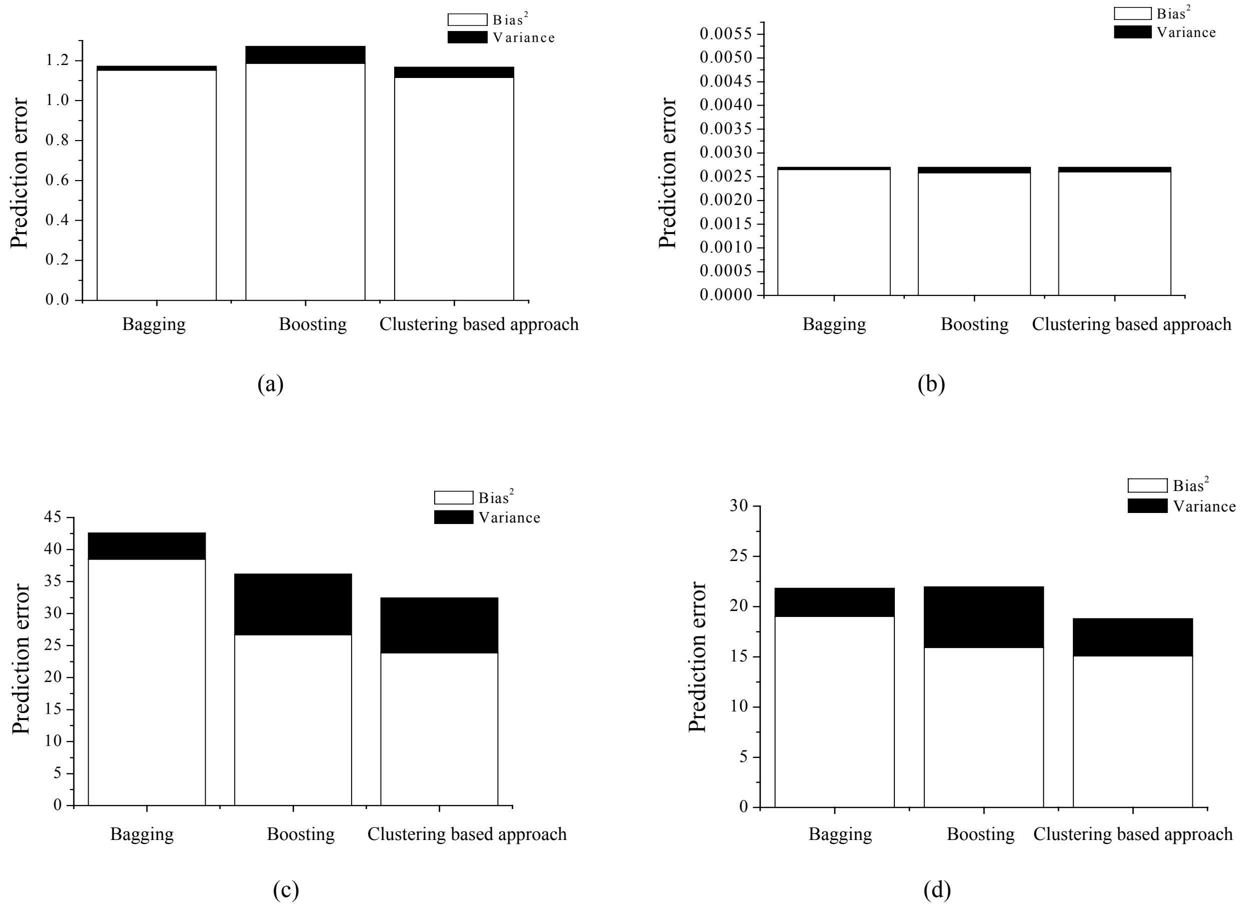
reduce the bias component will cause an increase in variance, and vice versa. Techniques in machine learning literature are often evaluated on how well they can optimize the trade-off between these two components.

Fig.2 reports the average result of bias-variance decomposition through 10-fold cross validation.

Previous works aver that Bagging works mainly through significantly reducing the variance, Boosting works through significantly reducing the bias (Bauer and Kohavi, 1999). Fig.2 shows that Bagging improves the predictive capability mainly by reducing variance and Boosting does it by reducing bias and variance. Fig.2 shows that Boosting is better than Bagging in reducing bias but that Bagging is better than Boosting in reducing variance. Compared with Bagging, clustering based approach slightly increases the variance but remarkably reduces bias. Compared with Boosting, clustering based approach can reduce the bias and variance at the same time. So we believe that the success of the clustering based approach may be able to its proper tuning of the bias/variance trade-off to reduce the prediction error (the sum of bias^2 and variance).

CONCLUSION

A selective algorithm based on clustering for constructing neural networks ensemble was proposed in this paper. Cluster technology was used to maintain the individual as diverse as possible. By comparison against other methods, we showed that this approach is effective and can generate far smaller ensemble with high performance. We also explained the mechanism of this approach by bias-variance decomposition. Further research and explore whether this approach can be extended to combine classifiers.



(a)

(b)

(c)

(d)

Fig.2 Bias-variance decomposition of three approaches
(a) Friedman #1 problem; (b) Plane problem; (c) Boston Housing problem; (d) Ozone problem

References

Bauer, E., Kohavi, R., 1999. An empirical comparison of voting classification algorithms: Bagging, Boosting, and variants. *Machine Learning*, **36**(1-2):105-139.

Blake, C., Keogh, E., Merz, C.J., 1998. UCI Repository of Machine Learning Databases. http://www.ics.uci.edu/~mlearn/MLRepository.html. Department of Information and Computer Science, University of California, Irvine, CA.

Breiman, L., 1996. Bagging predictors. *Machine Learning*, **24**(2):123-140.

Breiman, L., Friedman, J.H., 1985. Estimating optimal transformations in multiple regression and correlation (with discussion). *Journal of the American Statistical Association*, **80**:580-619.

Drucker, H., 1999. Boosting Using Neural Networks. *In*: Sharkey, A.(Ed.), Combining Artificial Neural Nets: Ensemble and Module Multi-net Systems. Springer-Verlag, London, p.42-49.

Friedman, J.H., Grosse, E., Stuetzle, W., 1983. Multidimensional additive Spline approximation. *SIAM Journal of Scientific and Statistical Computing*, **4**:292-301.

Fu, Q., Hu, S.X., Zhao, S.Y., 2004. A PSO-based approach for neural network ensemble. *Journal of Zhejiang University (Engineering Science)*, **38**(12):1596-1600 (in Chinese).

German, S., Bienenstock, E., Doursat, R., 1992. Neural networks and the bias/variance dilemma. *Neural Computation*, **4**(1):1-58.

Hansen, J.V., 2000. Combining Predictors: Meta Machine Learning Methods and Bias/variance and Ambiguity Decomposition. Ph. D Dissertation, Department of Computer Science, University of Aarhus, Denmark.

Hansen, L.K., Salamon, P., 1990. Neural network ensembles. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **12**(10):993-1001.

Krogh, A., Vedelsdy, J., 1995. Neural Network Ensembles Cross Validation, and Active Learning. *In*: Tesauro, G., Touretzky, D., Leen, T.(Eds.), Advances in Neural Information Processing Systems, Volume 7. MIT Press, Cambridge, MA, p.231-238.

Lazarevic, A., Obradovic, Z., 2001. Effective pruning of neural network classifier ensembles. *Proc. International Joint Conference on Neural Networks*, **2**:796-801.

Liu, Y., Yao, X., 2000. Evolutionary ensembles with negative correlation learning. *IEEE Trans. Evolutionary Computation*, **4**(4):380-387.

Melville, P., Mooney, R., 2003. Constructing Diverse Classifier Ensembles Using Artificial Training Examples. Proc. of the IJCAI-2003, Acapulco, Mexico, p.505-510.

Navone, H.D., Verdes P.F., Granitto, P.M., Ceccatto, H.A., 2000. Selecting Diverse Members of Neural Network Ensembles. Proc. 16th Brazilian Symposium on Neural Networks, p.255-260.

Opitz, D., Shavlik, J., 1996. Actively searching for an effective neural network ensemble. *Connection Science*, **8**(3-4):337-353.

Ridgeway, G., Madigan, D., Richardson, T., 1999. Boosting Methodology for Regression Problems. Proc. 7th Int. Workshop on Artificial Intelligence and Statistics. Fort Lauderdale, FL, p.152-161.

Rosen, B.E., 1996. Ensemble learning using decorrelated neural network. *Connection Science*, **8**(3-4):373-384.

Schapire, R.E., 1990. The strength of weak learn ability. *Machine Learning*, **5**(2):1971-227.

Zhou, Z.H., Wu, J.X., Jiang, Y., Chen, S.F., 2001. Genetic algorithm based selective neural network ensemble. *Proc. 17th International Joint Conference on Artificial Intelligence*, **2**:797-802.