ISSN 1392 – 124X INFORMATION TECHNOLOGY AND CONTROL, 2008, Vol.37, No.3

# ADAPTIVE COMMITTEES OF NEURAL CLASSIFIERS

**Arūnas Lipnickas**

*Kaunas University of Technology, Mechatronics Centre for Studies, Research and Information*
*Studentų St. 48-109, LT-51367 Kaunas, Lithuania*

**Abstract**. It is obvious that combination of several classifiers might improve overall classification performance. In this paper, on the contrary to the ordinary approach of utilising all neural networks available to make the committee decision, we propose to create adaptive committees, which are specific for each input data point. A prediction neural network is used to identify classifiers to be fused for making a committee decision about the given input data. The proposed technique is tested in three aggregation schemes and the effectiveness of the approach is demonstrated on the three real data sets.

**Keywords:** Adaptive committees, aggregation, neural networks, half&half sampling.

## 1. Introduction

It is well known that a combination of many different classifiers can improve classification accuracy. A variety of schemes have been proposed for combining multiple classifiers. The approaches used most often include the majority vote, the averaging, weighted averaging, the Bayesian approach, the Borda count, probabilistic aggregation, and aggregation by a neural network [1-3]. Often the researchers are focusing on the sophisticated combination methods and forgetting that the committee's performance is highly dependent on the members used. The committee members should be designed in the way by being accurate as well as diverse.

For some of the aforementioned approaches we can say that a combiner assigns weights of value to classifiers in one way or another. Aggregation schemes with the use of data-dependent weights, when properly estimated, provide higher classification accuracy [3, 5, 6].

The most predominant aggregation technique is by the use of all the networks available for making a committee decision. An alternative approach selects a single network, which is most likely to be correct for a given sample. In this case, aggregation weights are binary: $w_i \in \{0,1\}$, $i=1,\ldots,L$, where $L$ is the number of networks and $w_i = 1$ only for the most accurate network in the neighbourhood of a given sample. Thus only the output of the selected network is considered in the final decision.

In this paper, we propose an approach for building adaptive, data-dependent committees, which are specific for each input data point, in the way that, depending on an input data point, different set of classifiers is chosen to make a committee decision about the data point. A prediction neural network (NN) is trained to predict the behaviour of committee members for each data point from the training set, and is further used to select classifiers to be fused for making the committee decision (Figure 1, where $z_j$ and $p_j$ stand for the outputs of the networks).

As it is shown in Figure 1, the prediction neural network is used to identify classifiers to be fused for making a committee decision about a given input data. The $j^{th}$ output value of the predicting NN expresses the expectation level that the $j^{th}$ classifier will make a correct decision about the class label of a given input data.

Recently, it has been shown that half&half bagging through majority voting is capable to create very accurate committees of decision trees and neural networks [4, 5]. Data sampling by half&half bagging focuses on the most often misclassified data points from the training data set. We use the half&half sampling approach to create diverse neural network committees.

The proposed approach for adaptive classifiers selection is also compared with the committee design approach based on diversity measure named $\kappa$-*statistic* and approach with exhaustive search (ES) of members for the best performing committee.

Three real world problems are used to evaluate the approach proposed. We compare the developed technique with the ordinary decision fusion scheme when all the networks available are utilised to make a committee decision.

The paper is organised as follows. The neural network committee design approaches are briefly

described in the next section. The proposed approach for adaptive selection of committee members is presented in the third section. Section four describes the aggregation schemes. The databases used to test the approach proposed are briefly described in the section five. The sixth section presents the results of the experimental investigations. Finally, conclusions of the work are given in section seven.

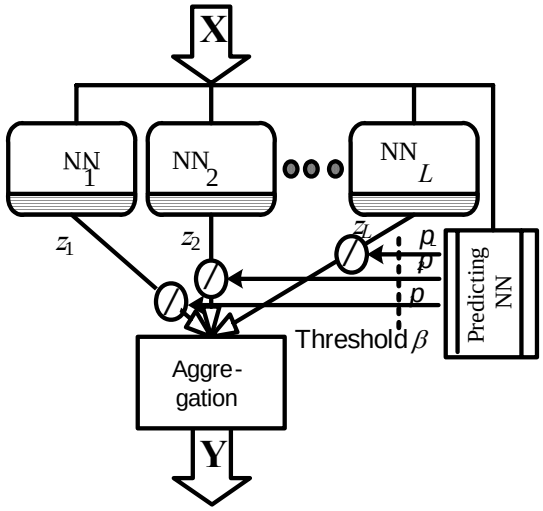


**Figure 1.** Architecture of the proposed combination scheme based on a dynamic neural network selection by a prediction NN neural network

## 2. Networks diversity and Half&Half sampling

Measuring the diversity of committee members is by no means trivial and there is trade-off between diversity and member accuracy. There are several approaches to measure members diversity [7], but unfortunately they work in a pairwise fashion. The result of a large set is the average of the pairwise measures for that set. One way to measure the diversity of neural networks is to calculate the $\kappa$-*statistic* as:

$$\kappa = \frac{\Theta_1 - \Theta_2}{1 - \Theta_2}, \qquad (1)$$

with $\Theta_1 = \sum_{i=1}^{Q} c_{ii} / N$ and $\Theta_2 = \sum_{i=1}^{Q} \left\{ \sum_{j=1}^{Q} \frac{c_{ij}}{N} \sum_{j=1}^{Q} \frac{c_{ji}}{N} \right\}$,

where $Q$ is the number of classes, $C$ is a $QxQ$ square matrix with $c_{ij}$ containing the number of data points assigned to class $i$ by the first network and into class $j$ by the second network and $N$ stands for the total number of used data.

The range for $\kappa$ values is from the interval 0 (*diverse*) to 1 (*correlated*), although they can be negative and range from -1 to 1. However, since there should be a positive correlation between the committee members, positive $\kappa$ values are expected. We used the statistic $\kappa$ to evaluate the diversities of trained neural networks committee.

*A. Half&Half sampling*

The basic idea of the half&half sampling is very simple. It is assumed that the training set contains $N$ data points. Suppose that $k$ classifiers have been already constructed. To obtain the next training set, randomly select a data point **x**. Present **x** to that subset of $k$ classifiers, which did not use **x** in their training sets. Use the majority vote to predict the classification result of **x** by the subset of classifiers. If **x** is misclassified, put it in set MC. If not, put **x** in set CC. Stop, when the sizes of both MC and CC are equal to M, where $2M \le N$. In [4], $M = N/4$ has been used. In this work, we investigate the effectiveness of the half&half sampling approach in creating accurate neural network committees for classification and compare it with two approaches: first with selecting strategy based on highest diversity of committee members and secondly by performing exhaustive search of the best possible combination from the set of available committee members.

*B. Committee design by $\kappa$ statistic*

As the comparison to half&half sampling approach, the $\kappa$ statistic has been used for selecting committee members with highest diversity. Obviously, $\kappa$ statistic works in a pairwise fashion and the result of larger set is the averaged values of the pairwise measures for that set. The minimum of the pairwise measure is always smaller or equal to the average value of the member set and is not suitable for selecting the optimal committee size. Hence, for this approach, the number of the classifiers to be used is always determined in advance.

*C. Committee selection by exhaustive search*

For the comparison to the two aforementioned approaches the exhaustive search (ES) procedure on training data has been performed for finding the committee members with the best overall performance. It should be noted that for the exhaustive search procedure all possible combinations have to be evaluated. In the case of $L$ classifiers the number of considered combinations is $2^L$. For reducing the computational burden the data points, where all classifiers agree, were removed from the training data set. Clearly, there are no advantages in combining the identical networks, no matter how ingenious a combination method is employed.

## 3. Proposed approach for adaptive committees

As it is shown in Figure 1, the prediction neural network is used to identify classifiers to be fused for making a committee decision about the given input data. The $j^{th}$ output value of the prediction network expresses the expectation level that the $j^{th}$ classifier will make a correct decision about the class label of the given input data. The networks whose probabilities to

be accurate are higher than threshold $\beta$, are afterwards involved in decision aggregation process.

A prediction network is trained in the manner to predict the behaviour of committee members for each data point from the training set. The procedure for training data collection of predicting neural network and determination of threshold $\beta$ is explained below.

The procedure is encapsulated in the six steps:

1. Divide the available data into training, test, and cross-validation data sets.
2. Train $L$ neural networks using the half&half sampling technique.
3. Classify the training data set by all networks of the committee.
4. For each training data vector $\mathbf{x}_i$, form a $L$-dimensional target vector $\mathbf{t}_i = [t_{i1},\ldots, t_{iL}]^T$, with $t_{ij}$=1 if the $\mathbf{x}_i$ data vector was correctly classified by the $j^{\text{th}}$ network, and $t_{ij}$=0, otherwise.
5. Using the training data set and the target vectors obtained in Step 4, train a neural network to predict whether or not the classification result obtained from the $L$ networks for an input data point $\mathbf{x}$ will be correct. The prediction network consists of $L$ output nodes and $n$ input nodes, where $n$ is the number of components in $\mathbf{x}$. Therefore, each output node stands for one particular network. The number of hidden nodes needs to be determined.
6. Determine the optimal threshold value $\beta$ for including neural networks into a committee. The $j^{\text{th}}$ network is included into a committee if $p_j>\beta$, where $p_j$ is the $j^{\text{th}}$ output of the prediction network. The value $\beta$ is the value yielding the minimum cross-validation data set classification error obtained from a committee of the selected networks.

Having the threshold $\beta$ determined, *data classification* proceeds as follows:

1. Present a test data point $\mathbf{x}$ to the prediction network and calculate the output vector $\mathbf{p}=[p_1,\ldots,p_L]$.
2. Classify the data point by the networks satisfying the condition $p_j>\beta$.
3. Aggregate the outputs of the selected networks into a committee decision according to a chosen combination algorithm.

3.1. If condition "2" rejects all members, then classify data with network $i = \arg\max_{j=1,\ldots,L}\left(p_j\right)$.

Note that the optimal threshold value is determined in the training phase and then fixed for the use in the classification phase. Also note that the build committee is specific for each input data point. This seems reasonable, since the $L$ neural networks may have different accuracy in different regions of the input space.

We investigate three schemes for aggregating the outputs of the adaptively selected networks. In the context of the aggregation schemes used, we compare the proposed concept with an ordinary decision aggregation approach, when all the trained networks are utilised to make a committee decision.

## 4. Aggregation schemes used

To test the proposed approach, we used three simple aggregation schemes that do not utilise any aggregation parameters, namely the *majority vote, averaging*, and the *median* aggregation rule. We now briefly describe the aggregation schemes used.

**Majority vote**. The correct class is the one chosen by the most neural networks. If all the neural networks indicate different classes, then the neural network with the overall maximum output value is selected to indicate the correct class. Ties can be broken in various ways; one of them is to avoid the even number of committee size.

In our case, if even number of committee members is selected then an additional member from the rest of classifiers with highest probability value $p_j$ is included into committee decision.

**Averaging**. This approach simply averages the individual neural network outputs. The output yielding the maximum of the averaged values is chosen as the correct class $q$:

$$q = \arg\max_{j=1,\ldots,Q}\left(Z_j = \frac{1}{L}\sum_{i=1}^{L} z_{ji}(\mathbf{x})\right), \qquad (2)$$

where $Q$ is the number of classes, $L$ is the number of neural networks, $z_{ij}$ $(\mathbf{x})$ represents the $j^{\text{th}}$ output of the $i^{\text{th}}$ network given an input pattern $\mathbf{x}$, and $Z_j(\mathbf{x})$ is the $j^{\text{th}}$ output of the committee given an input pattern $\mathbf{x}$.

**Median rule**. In some cases, when a classifier in a combined group is very sensitive to outliers, then the group decision could lead to an error. It is well known that a robust estimate of the averaging is the median. The median combination leads to the following rule:

$$q = \arg\max_{j=1,\ldots,Q}\left(Z_j = \operatorname*{med}_{i=1}^{L}\left(z_{ji}(\mathbf{x})\right)\right). \qquad (3)$$

## 5. Experimental testing

For the experimentation the three databases were selected. From the ELENA project we have chosen the two real data sets, *Phoneme* (2 classes, 5 features and 5404 samples) and *Satimage* (6 classes, 5 features and 6435 samples). The additional, *Thyroid* database (3 classes, 21 features and 7200 samples), has been taken from a collection called PROBEN 1, which represents a medical diagnosis task.

All comparisons between the different aggregation schemes presented here have been performed by leaving aside 10% of the data available as a *Cross-Validation* data set and then dividing the rest of the

data into *Training* and *Test* sets of equal size. In all the tests, one hidden layer MLPs with 10 sigmoidal hidden units served as committee members. This architecture was adopted after some experiments. Since we only investigate aggregation approaches, we have not performed expensive experiments for finding the optimal network size for each data set used. We run each experiment ten times, and the *mean* errors and *standard deviations* of the errors are calculated from these ten trials. In each trial, the data set used is randomly divided into *Training, Cross-Validation,* and *Test* parts.

In the first set of experiments, we investigated the ability of the half&half sampling technique to create diverse and accurate neural networks. The size of the committees was grown from 3 to 20 members. Afterwards the committee design approach based on $\kappa$ statistic has been involved to select diverse committee members from the total of 20 members. The selection criterion was the lowest averaged value of pairwise diversity measure. Finally, for comparison purpose the procedure for exhaustive search was involved for finding optimal committees with varied size from 3 to 20 members.

Figure 2, curves with squares, crosses and pentagrams, illustrates the *Test* set classification error of the committees for the different databases as a function of the committee size. Aggregation by the majority vote rule has been used in these experiments. The notation "H&H" stands for the half&half sampling approach, "ES" – for exhaustive search (ES) procedure and "$\kappa$ stat" for selection approach based on $\kappa$ statistic.

As can be seen from Figure 2, the half&half sampling approach performs similar to exhaustive search (ES) procedure and outperforms the approach based on highest diversity measure. The lower classification error is faster reached by ES approach and with smaller committee size. The averaged diversity measure of the 20 committee members for the databases *Phoneme, Satimage* and *Thyroid* are 0.46; 0.81 and 0.75, respectively. The averaged values of diversity measure as a function of the committee size for the *Satimage* data set are shown in Fig. 3. The selection approach based on $\kappa$ statistic tries to keep diverse committees neglecting the individual performance of members. This criterion does not take into account the difference in errors and correct classification results; obviously, it is beneficial when members agree on correct result and disagree on misclassifications. Contrary, the H&H approach selects much more accurate committee members.

The similar pattern of accuracy and diversity was observed across the other data sets. We can, therefore, conclude that the half&half sampling technique is capable to create diverse and sufficiently accurate neural networks. Its performance is only slightly worse than that of the exhaustive search approach. Other comparisons and evaluations of the half&half sampled committees can be found in [4, 5].

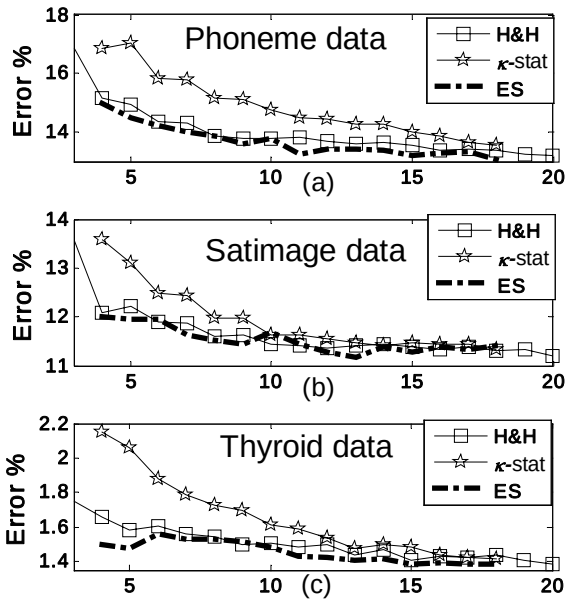


**Figure 2.** Classification error as a function of the committee size for: a) the *Phoneme* data set, b) the *Satimage* data set, c) the *Thyroid* data set

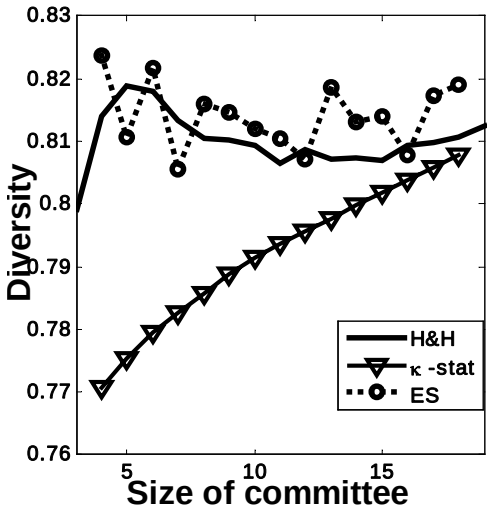


**Figure 3.** Diversity measure as a function of the committee size for the *Satimage* data set

In the next set of experiments, we investigated the effectiveness of the adaptive networks selection technique in creating accurate neural network committees. The regularised training techniques have been employed in training neural classifiers. In the ordinary decision aggregation approach, without the proposed neural networks selection procedure, we utilised committees consisting of 20 members. The actual average size of the committees created by the proposed procedure was considerably smaller. The prediction network was found with having 15 nodes in the hidden layer. Tables 1 and 2 summarize the *Test* data set classification error obtained in these tests. The following notations are used in the tables: *Mean* stands for the percentage of the average test set classification error, *Std* is the standard deviation of the error, and The best

means the single neural network with the best average performance.

As can be seen from the tables, there is an obvious improvement in classification accuracy when combining networks. All the three aggregation schemes yielded approximately the same performance. The approach with the proposed neural networks selection technique is slightly superior to the usual aggregation approach when all the networks available are aggregated to make a committee decision.

Table 2 provides the minimal, average and maximal numbers of neural networks included into a committee from the 20 available for the aggregation by majority voting rule and databases used. The table also presents minimal, average and maximal values of the selection threshold $\beta$ found for the different cases. The value of $\beta = 0$ implies using all the networks available to make committee decisions. The average number of selected neural networks is far below the 20 available. Therefore, the proposed technique allows reducing both classification error and computational time by removing unreliable classifiers.

**Table 1.** The test data set classification error rate obtained from the half&half sampled neural network committees fused by the Majority Vote, Averaging, and Median aggregation rules

| Database | The best | | Majority | | Averaging | | Median | |
|---|---|---|---|---|---|---|---|---|
| | Mean | Std | Mean | Std | Mean | Std | Mean | Std |
| **Without selection** | | | | | | | | |
| *Phoneme* | 15.80 | 0.6 | 13.20 | 0.6 | 13.20 | 0.6 | 13.10 | 0.5 |
| *Satiamge* | 12.80 | 0.5 | 11.20 | 0.6 | 11.10 | 0.4 | 11.30 | 0.4 |
| *Thyroid* | 2.08 | 0.4 | 1.38 | 0.16 | 1.38 | 0.12 | 1.40 | 0.12 |
| **With proposed selection** | | | | | | | | |
| *Phoneme* | 15.80 | 0.6 | 12.80 | 0.6 | 12.80 | 0.4 | 12.90 | 0.4 |
| *Satiamge* | 12.80 | 0.5 | 11.19 | 0.3 | 11.05 | 0.4 | 11.17 | 0.3 |
| *Thyroid* | 2.08 | 0.4 | 1.27 | 0.13 | 1.25 | 0.15 | 1.29 | 0.13 |

Figure 4 plots the *Test* data set classification error rate and averaged size of the committees for the *Phoneme* data set as a function of the neural network selection threshold $\beta$. In this experiment, the majority vote rule has been used to aggregate the selected networks into a committee. Figure 4 shows the strong dependence between the threshold value, the averaged size of committee and the classification error rate. As it can be seen from Figure 4, by increasing threshold $\beta$ the number of selected neural networks decreases, but committee error increases. It is the trade off between accuracy and the size of the committee.

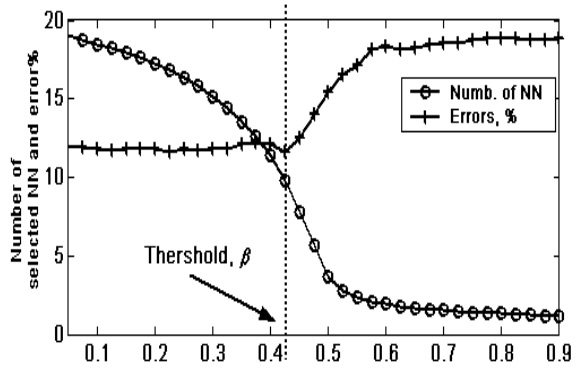


**Figure 4.** The test data set classification error rate and averaged number of selected neural networks of the committee for the *Phoneme* data set as a function of the selection threshold $\beta$

**Table 2.** The average number of selected neural networks from the 20 available and the average value of the optimal selection threshold found for the aggregation by majority vote rule

| | half&half sampling | |
|---|---|---|
| Database | # Selected NN | Threshold $\beta$ |
| *Phoneme* | 9< 14 <17 | 0.1< 0.25< 0.45 |
| *Satimage* | 12< 14 <17 | 0.2< 0.26< 0.35 |
| *Thyroid* | 13< 17 <19 | 0.7< 0.8< 0.95 |

## 6. Conclusions

In this paper, we used the half&half sampling technique to collect data sets for training neural network committees. In all the tests performed, the half&half data sampling approach outperformed the committee member's selection approach based on $\kappa$ statistic and it was only slightly worse that ES approach. Obviously, the pairwise selection criterion cannot guarantee the optimal committee. Good committees should be diverse as well as accurate. The exhaustive search procedure can guarantee the best committee only after exhausting searching procedure.

An approach to create adaptive committees of neural network for classification from already trained pool of classifiers was proposed. The approach banks on the idea of having a specific committee for each input data point. Different networks and a different number of them may be adaptively selected and fused

into a committee to make a decision about different input data points. The networks utilised are determined by those outputs of a prediction network, the output value at which exceeds a particular selection threshold. The $j^{th}$ output value expresses the expectation level that the $j^{th}$ classification neural network will make a correct decision about the class label of a given input data point.

The effectiveness of the proposed approach in creating accurate neural network committees for classification was investigated using three real data sets. The proposed approach was compared with the scheme of the ordinary neural networks fusion. The comparisons were made for three neural networks aggregation approaches, named *majority vote*, *averaging*, and aggregation by the *median rule*. In all the tests performed, the proposed way of generating neural network committees was superior to the ordinary decision fusion scheme when all the networks available are utilised to make a committee decision.

## References


**[1]** **A. Verikas et al.** Soft combination of neural classifiers: A comparative study. *Pattern Recognition Letters*, 1999, *Vol.*20, 429-444.

**[2]** **A. Verikas, A. Lipnickas, M. Bačauskienė, K. Malmqvist.** Fusing neural networks through fuzzy integration. *In H. Bunke, A. Kandel, editors, Hybrid Methods in Pattern Recognition, World Scientific*, 2002, 227-252.

**[3]** **A. Verikas, A. Lipnickas.** Fusing neural networks through space partitioning and fuzzy integration. *Neural Processing Letters, Kluwer Academic Publishers*, 2002, *Vol.*16, *No.*1, 53-65.

**[4]** **L. Breiman.** Half&Half bagging and hard boundary points. *Technical report* 534, *Statistics Departament, University of California, Berkeley*, 1998. *www.stat.berkley.edu /users/breiman*.

**[5]** **A.Lipnickas, J. Korbicz.** Adaptive Selection of Neural Networks for a Committee Decision. *Proceedings of the second IEEE international workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications* (IDAACS'2003), 2003, 109-114.

**[6]** **A. Lipnickas.** Classifiers Fusion With Data Dependent Aggregation Schemes. *Proceedings of the* 7*th International Conference on Information Networks, System and Technologies*, ICINASTe'2001, *Minsk, Belarus, October* 2-4, 2001, 147-154.

**[7]** **L.I. Kuncheva, C.J. Whitaker.** Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 2003, *Vol.*51. 181-207.


Received April 2008.