

# Chapter 8

## Conclusions and Future Work

In this thesis, I have presented a systematic study of neuro ensembles on binary classification problems. First, I have investigated the relationship between diversity and accuracy of a number of state-of-the-art neuro ensembles, aiming to generate insights into how to design a good neuro ensemble. Applying these insights, I have proposed a method to co-evolve the gate and experts of a mixture of experts (ME) using a cooperative coevolution (CC) framework, which I name the cooperative coevolutionary mixture of experts or CCME. The proposed method blends concepts from mixture of experts and cooperative coevolution to generate a modularized neuro ensemble, which solves both the diversity problem - by forcing the experts to work on different regions of input (ME) and to be diverse in terms of species diversity (CC) - and the generalization problem - the whole system must work cooperatively. The more conventional ME is then compared against CCME in terms of the generalization performance and running time. I have also investigated and presented the traditional back propagation ME and the proposed method in the light of automatic problem decomposition, using a newly-derived set of visualization tools. Finally, I have shown the effects of regularization, especially the so-called learning by forgetting, on the structural complexity and the behaviors of ME and CCME.

The main findings from the research work in this thesis can be summarized as

follow:

1. This thesis has investigated different aspects of a number of state-of-the-art neuro ensemble methods (Chapter 4): namely the Simple Ensemble, the Negative Correlation Learning Ensemble, the Island Ensemble and the Multi Objective Ensemble.

The results verify a number of points raised in the literature:

- (a) Combining individual networks into an ensemble improves the performance of the system.
  - (b) Different combination gates have little effect on the average performance of the ensemble
  - (c) The diversity level maintained by negative correlation learning is poor.
  - (d) Local search helps evolution to find better solutions
  - (e) Noise injection shows interesting effects on performance enhancement, though the improvement is not yet clear.
  - (f) Early stopping enhances generalization
  - (g) The connection between diversity and performance of the ensemble remains a hypothesis with limited or no verification
2. The proposed Cooperative Coevolutionary Mixture of Experts (CCME) method is validated against a set of benchmark binary classification problems. I have analyzed different aspects of both the traditional back propagation ME and the novel CCME in chapter 5. The key findings are:

- (a) In terms of performance, CCME is better on average than the traditional ME in classification problems.
- (b) CCME is comparable to ME in terms of the running time.
- (c) Early stopping is useful in enhancing the generalization of both back propagation ME and CCME.

- (d) CCME and ME are robust to different error functions, learning rates, number of experts and network complexity in a number of binary classification problems.
3. The proposed visualization tools are applied to analyze the problem decomposition behaviors of ME and CCME (Chapter 6). The key results are:
- (a) Both ME and CCME can decompose the input space into less complex regions (i.e. sub-tasks) in such a way that the available experts are able to classify the data in their clusters with higher accuracy.
  - (b) The tools show how ME and CCME can discover the modularity of the problem if there is any.
  - (c) Increasing network complexity, in terms of number of hidden units, helps the experts to classify their clusters with better accuracy. However, there is a phase transition, beyond which an increase in complexity deteriorates the system performance.
  - (d) Increasing the number of experts also enhances the system performance by dividing the data into more clusters. Again, however, there is a critical region, beyond which adding more experts does not improve the system performance.
4. The thesis has extended both ME and CCME models by adding a regularization term - based on learning by forgetting and weight elimination - during training (Chapter 7). Another contribution of this chapter is that it introduces and uses a number of novel visualization tools to visualize the weight distribution and the architecture of the model. The key results are:
- (a) In terms of accuracy, regularization does not significantly affect the performance of the models.
  - (b) In terms of network complexity, learning by forgetting is effective in pushing irrelevant weights toward zero, while maintaining the same level of accuracy.

- (c) weight elimination is not as effective as learning by forgetting in pruning out the irrelevant weights
- (d) Because of its structural modularization ability (Ishikawa 1996; Ishikawa and Yoshino 1993), learning by forgetting can manipulate the models into doing different types of tasks. In some cases, it can turn the ME model into a traditional clustering system by pruning all the input-to-hidden connections of the experts, and letting the gate divide the data into clusters of the same class. Another benefit of learning by forgetting is to allow different experts of the models to select different, but not necessarily disjoint, sets of input features.
- (e) There is a possible relationship between the weight distribution, in terms of magnitude, and the generalization ability of the ME and CCME model.

In conclusion, the experiments and findings answer the research question: artificial evolution can produce neuro ensembles that automatically decompose complex classification problems by dividing the data to sub-regions where the input-output relationship is easier to learn and assigning different experts to these sub-regions. The higher performance of CCME over the conventional ME implies that by adding a cooperative co-evolution layer, the system can explore the search space more efficiently (CCME outperforms the random search for ME) and thus find more fitting neuro ensembles.

## 8.1 Future Work

Numerous directions for further explorations and investigations have emerged from the work of this thesis. Some open research questions have already been highlighted in the respective chapters where they directly arise from the experiments and analysis. These can be summarized as follows:

1. In chapter 6, the automatic problem decomposition analysis of CCME shows that the model can find a very good way to divide the data into better clusters, where

“better” measures the ease with which an expert can classify the data in the clusters. The results suggest the following questions: (1) can we define a measure for the ease with which a cluster may be classified?, and if yes, (2) can this measurement be used to guide the model toward dividing the data in a better way, or can it at least explain the differences in performance between methods of automatic problem decomposition?

2. In section 7.3.2.4.1, the correlation between the generalization error curves and the weight distribution plots suggests the following open research directions: (1) is there a relationship between the weight distribution and the generalization error?, (2) in the case of regularization, can the phase changes in the weight distribution be used as an estimate for the phase change in the generalization errors? and (3) if there is such a relationship, can one find an optimized weight distribution which will produce an optimized neuro ensemble in terms of the generalization ability?

Besides the above questions, a number of research directions arising from the philosophical issues underlying the thesis are outlined here:

First, a critical question, that still remains open in the ensemble literature, is how to decide on the optimum ensemble size. As I discussed in the thesis, cooperative coevolution is a powerful framework allowing a suitable number of sub-populations to emerge, based on the fitness of the whole system. The principle is to add and remove sub-populations when the overall fitness stagnates for a number of generations (Potter 1997; Potter and De Jong 2000). The proposed framework allows the ensemble size to emerge as the system adjusts itself to the best fitness.

Second, in this thesis, I have shown how ME and CCME automatically decompose a hard problem into sub-regions of the input with the property that the sub-regions are easier, in terms of input-output relationship, for the available classifiers to solve. However, the decomposition mechanism underlying the models remains an open question. As suggested, a measure of the ease-of-classification of the sub-regions could be

very valuable in guiding the system towards better decomposition schemes. If such a measure exists, it could be cooperated into the performance fitness of the system, for example serving as an objective in a multi-objective optimization method. It would force the gate to divide the data into optimum regions, in terms of simplest input-output relationship, and at the same time evolve accurate experts to solve these sub-regions with minimum error.

Finally, although learning by forgetting is applied and analyzed in this thesis, it is applied in a simple form in which the regularization parameters are the same for both the gate and the experts. As suggested in the experiments and analysis in chapter 7, the relationship between these regularization parameters and the weight distributions of the gate and the experts is not simple. It is therefore interesting to further investigate the effect of LF with different parameters for these components. Also, it will be valuable to study the relationship between the weight distribution and the error rate, since such relationship might give valuable insights into the optimum range and shape of the weight distribution to generate neuro ensembles that generalize well. If such an optimum weight distribution can be found, then a mixture of experts can be quickly designed to solve the problem. Besides learning by forgetting, there are other successful regularization schemes in the literature of ANN. It would be interesting to study the effects of these schemes on both ME and CCME.