Generalization and Catastrophic Forgetting in Radial Basis Function Networks

Neil Middleton
Department of Computer Science
Brunel University, Uxbridge
West London, UB8 3PH, UK.
Neil.Middleton@brunel.ac.uk

Connectionist models

Connectionist networks have been widely used to model human categorization. To be psychologically plausible such networks must at least be able to generalize, and not be prone to catastrophic forgetting.

Multi Layer Perceptrons

The multi layer perceptron (MLP) forms distributed hidden layer representations, and stores training patterns in superpositional form. Potentially, all the network resources (weights and processing units) are used to represent each learned pattern. This can result in catastrophic interference in which learning a single new pattern causes the network to unlearn (or forget) all previously learned patterns.

MLPs learn classification tasks by fitting the category boundaries, and so are able to generalize to treat unseen patterns in a reasonable way.

Radial Basis Function Networks

Radial basis function (RBF) networks were originally presented as localist networks with one hidden unit to store each training example. Such networks make good models of exemplar theories of categorization (Kruschke, 1992), and the orthogonality of hidden layer representations prevents catastrophic forgetting. However, localist networks often do not generalize well.

How existing models work

Feed forward connectionist networks learn nonlinearly separable tasks by recoding input vectors as linearly separable vectors in hidden layer activation space. Orthogonal vectors are linearly separable, so a suitable recoding is to use one hidden unit to store each training pattern, as in the original RBF network. Linearly separable patterns are typically not orthogonal, so extended and distributed hidden layer representations can also be used to learn the task, as in MLPs.

Generalization without forgetting

French (1994) argued that to prevent catastrophic interference in MLPs while retaining the ability to generalize, hidden layer activation vectors must be as orthogonal and as distributed as possible. French presents an algorithm designed to achieve this trade off by decreasing the extendedness of hidden layer representations in MLPs.

An alternative approach is to start with a localist RBF network and allow the network to form more extended and distributed representations while retaining as much orthogonality as possible. This can be achieved by making the receptive fields of RBF hidden units larger. The network remains localist in the sense that the hidden units store an (interpretable) pattern, yet is distributed since each unit plays a part in representing more than one pattern.

The network can also be forced to create less localist representations if it is given fewer hidden units than there are training patterns.

Experiments have been performed with RBF networks for learning simple pattern classification tasks. These experiments show how RBF networks are able to generalize and at the same time form relatively orthogonal hidden layer representations.

Conclusion

It has been argued that the properties of RBF networks for cognitive modelling have still to be fully investigated.

As further work, the individual properties of MLPs and RBF networks could be utilized in a single model. In performing classification tasks RBF networks fit the training examples and MLPs fit category boundaries. This suggests that a hybrid RBF and MLP network could be used to model both boundary reference point effects and exemplar theories of human categorization.

Acknowledgements

I acknowledge the support of the Engineering and Physical Sciences Research Council through a PhD studentship.

References

French, R.M. (1994). Dynamically constraining connectionist networks to produce distributed, orthogonal representations to reduce catastrophic interference. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 335-340). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kruschke, J.K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99, 22-44.