Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Neurocomputing 70 (2006) 155–163
NEUROCOMPUTING
www.elsevier.com/locate/neucom

A novel mixture of experts model based on cooperative coevolution

Minh Ha Nguyen[a,*], Hussein A. Abbass[a], Robert I. Mckay[b]

[a]*Artificial Life and Adaptive Robotics (A.L.A.R.) Lab, The ARC Centre for Complex Systems, School of Information Technology and Electrical Engineering, Australian Defence Force Academy, University of New South Wales, Canberra, ACT 2600, Australia*

[b]*School of Computer Science and Engineering, Seoul National University San 56-1, Sinlim-dong, Gwanak-gu, Seoul 151-744, Republic of Korea*

Received 8 December 2005; received in revised form 6 February 2006; accepted 17 April 2006
Communicated by K. Li
Available online 4 August 2006

Abstract

Combining several suitable neural networks can enhance the generalization performance of the group when compared to a single network alone. However, it remains a largely open question, how best to build a suitable combination of individuals. Jacobs and his colleagues proposed the mixture of experts (ME) model, in which a set of neural networks are trained together with a gate network. This tight coupling mechanism enables the system to (i) encourage diversity between the individual neural networks by specializing them in different regions of the input space and (ii) allow for a "good" combination weights of the ensemble members to emerge by training the gate, which computes the dynamic weights together with the classifiers.

In this paper, we have wrapped a cooperative coevolutionary (CC) algorithm around the basic ME model. This CC layer allows better exploration of the weight space, and hence, an ensemble with better performance. The results show that CCME is better on average than the original ME on a number of classification problems. We have also introduced a novel mechanism for visualizing the modular structures that emerged from the model.

Keywords: Neuro-ensemble; Artificial neural networks; Mixture of experts; Cooperative coevolution

1. Introduction

Evolutionary artificial neural networks (EANNs) have been widely studied in the last few decades. The main power of artificial neural networks (ANNs) lies in their ability to correctly learn the underlying function or distribution in a data set from a sample. This ability is called generalization. Mathematically, the generalization ability can be expressed in terms of minimizing the recognition error of the neural network, on previously unseen data. Thus evolutionary computation (EC), a global optimization approach, can be employed to optimize this error function. As discussed in the prominent review of Yao [31], evolutionary methods can be applied on different levels of ANN, such as the architecture and the connection weights.

Much of the ANN literature concentrates on finding a single solution (network) to learn a task. However, an optimum network on the training data (i.e. seen data) may not generalize well on the testing data (i.e. unseen data). An ANN could either overtrain/overfit (memorizing the data rather than learning the correct distribution) or undertrain (not trained enough, or too simple, to fit the data well) (see [7] on the bias/variance dilemma and the generalization problem). Many published works have shown that an ensemble of neural networks (i.e. neuro-ensemble) can generalize better than individual networks [10,16–18,23,25,32–34]. The main argument for neuro-ensembles is that different members of the ensemble may possess different bias/variance trade-offs, hence a suitable combination of these biases/variances could result in an improvement in the generalization ability of the whole ensemble [25,34]. It is obvious that a self-similar set of

*Corresponding author.
E-mail addresses: minhha_76@yahoo.com (M.H. Nguyen), h.abbass@adfa.edu.au (H.A. Abbass), rim@cse.snu.ac.kr (R.I. Mckay).

0925-2312/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2006.04.009

individuals is not desirable, since it multiplies the effort to train them without adding to the overall performance—i.e. the system's performance is similar to that of a single network.

One important application of neuro-ensembles is in problem decomposition. Most real-world problems are too complicated for a single individual to solve. Divide-and-conquer has proved to be efficient in many of these complex situations. The issues are (i) how to divide the problem into simpler tasks, (ii) how to assign individuals to solve these subtasks and (iii) how to synthesize the whole system back together. If the problem has a distinct natural decomposition, it would be possible to derive such a decomposition by hand. However, in most real-world problems, we either know too little about the problem, or it is too complex for us to have a clear understanding on how to hand-decompose it into subproblems. Thus, it is desirable to have a method to automatically decompose a complex problem into a set of overlapping or disjoint subproblems, and to assign one or more specialized problem solving tools or experts to each of these subproblems. The remaining question is how to combine the outputs of these experts if the decomposition scheme is unknown in advance.

Jacobs [8,9] has proposed an ensemble method called mixture of experts (ME), based on the divide-and-conquer principle. In their method, instead of assigning a set of combinational weights to the experts, an extra gating component is used to compute these weights dynamically from the inputs (Fig. 1). This gating component is trained, together with other experts, through a specially tailored error function, which localizes the experts into different subsets of the data while improving the system's performance. In the ME model, the expert could be of any type, e.g. an ANN or a C4.5 decision tree, but the gating is often an ANN. Jordan and Jacobs [11,12] extended the model to the so-called hierarchical mixture of experts (HME), in which each component of the ME model is replaced with an ME model. Since Jacobs' proposal of the ME model in 1991, there has been a wide range of research into it.

Some authors [1,13,14] have established how the ME model works in statistical terms. Waterhouse [28,30] and Moeland [19] have applied the Bayesian framework to design and explain the ME model. According to their interpretation, the ME output(s) can be considered as estimates of the posterior probabilities of class membership [19]. Thus, the Bayesian framework can be used to design the training error function [3] and estimate the parameters for the ME model [30]. Besides the original ME model, a large number of variants have been put forward. Waterhouse and Cook [29] and Avnimelech and Cook [2] proposed to combine ME with the boosting algorithm. They argued that, since boosting encourages classifiers to be experts on different patterns that previous experts disagree on, it can split the data set into regions for the experts in the ME model, and thus ensure localization of experts. The dynamic gating function of the ME ensures a good combination of classifiers [2]. Tang et al. [26] tried to explicitly localize the experts by applying a self-organizing map to partition the input space for the experts. Wan and Bone [27] used a mixture of radial basis function networks to partition the input space into statistically correlated regions and learn the local covariation model of the data in each region.

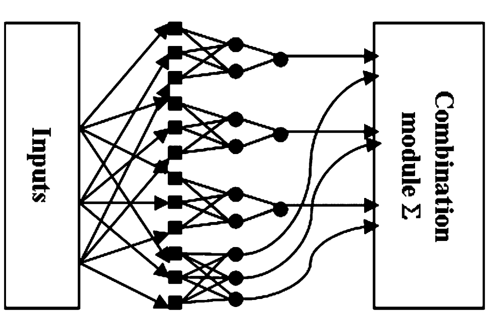

Fig. 1. Mixture of expert architecture.

Although gradient descent is the most popular ANN training method, especially in industrial problems, because of its simple implementation and its efficiency, it has some serious drawbacks. The growing literature on EC research as a global optimization method led to a number of successful attempts to evolve ANNs [31]. A newer branch of EC called the cooperative coevolutionary (CC) algorithm was proposed by Potter and De Jong [21,22].

Garcia-Pedrajas et al. [6] have applied multiobjective optimization in conjunction with CC, to evolve a set of subpopulations of well-performed, regularized, cooperative and diverse ANNs which can be used in a set of ensembles. Khare et al. [15] used the concept of CC on a set of subpopulations of radial basis function networks, where each subpopulation is designed to solve a particular subtask of the whole problem. A second level, consisting of a swarm of ensembles, to combine selected ANNs from the subpopulation, is also evolved in parallel with these subpopulations. The two disadvantages of this method are (i) its requirement for a prior knowledge about the problem in order to know the fixed number of required modules and (ii) its dependence on credit assignment, in that the fitness of each module is decided by the contribution of the module to the whole system. To solve the problem of fixed number of modules, Khare et al. [15] suggested using Potter's approach of adding and removing subpopulations whenever the system's fitness stagnates for a predetermined period. Despite the remaining credit assignment problem, their method has the merit that both the structures and parameters, of both the modules and the whole ensemble, can be evolved within the framework.

2. Mixture of experts

The ME model consists of a number of experts combined through a gate (Fig. 1), all having access to the input space. The components can be any type of classifiers—in this paper, we use simple feed-forward multilayer neural