Ensemble Learning for Regression

Niall Rooney
University of Ulster, UK

David Patterson
University of Ulster, UK

Chris Nugent
University of Ulster, UK

INTRODUCTION

The concept of ensemble learning has its origins in research from the late 1980s/early 1990s into combining a number of artificial neural networks (ANNs) models for regression tasks. Ensemble learning is now a widely deployed and researched topic within the area of machine learning and data mining. Ensemble learning, as a general definition, refers to the concept of being able to apply more than one learning model to a particular machine learning problem using some method of integration. The desired goal of course is that the ensemble as a unit will outperform any of its individual members for the given learning task. Ensemble learning has been extended to cover other learning tasks such as classification (refer to Kuncheva, 2004 for a detailed overview of this area), online learning (Fern & Givan, 2003) and clustering (Strehl & Ghosh, 2003). The focus of this article is to review ensemble learning with respect to regression, where by regression, we refer to the supervised learning task of creating a model that relates a continuous output variable to a vector of input variables.

BACKGROUND

Ensemble learning consists of two issues that need to be addressed, *ensemble generation*: how does one generate the base models/members of the ensemble and how large should the ensemble size be and *ensemble integration:* how does one integrate the base models' predictions to improve performance? Some ensemble schemes address these issues separately, others such as Bagging (Breiman, 1996a) and Boosting (Freund & Schapire, 1996) do not. The problem of ensemble generation where each base learning model uses the same learning algorithm (*homogeneous learning*) is generally addressed by a number of different techniques: using different samples of the training data or feature subsets for each base model or alternatively, if the learning method has a set of learning parameters, these may be adjusted to have different values for each of the base models. An alternative generation approach is to build the models from a set of different learning algorithms (*heterogeneous learning*). There has been less research in this latter area due to the increased complexity of effectively combining models derived from different algorithms. Ensemble integration can be addressed by either one of two mechanisms: either the predictions of the base models are combined in some fashion during the application phase to give an ensemble prediction (*combination approach*) or the prediction of one base model is selected according to some criteria to form the final prediction (*selection approach*). Both selection and combination can be either *static* in approach, where the learned model does not alter, or *dynamic* in approach, where the prediction strategy is adjusted for each test instance.

Theoretical and empirical work has shown that if an ensemble technique is to be effective, it is important that the base learning models are sufficiently *accurate* and *diverse* in their predictions (Hansen & Salomon, 1990; Sharkey, 1999; Dietterich, 2000). For regression problems, accuracy is usually measured based on the training error of the ensemble members and diversity is measured based on the concept of ambiguity or variance of the ensemble members' predictions (Krogh & Vedelsby, 1995). A well known technique to analyze the nature of supervised learning methods is based on the bias-variance decomposition of the expected error for a given target instance (Geman et al., 1992). In effect, the expected error can be represented by three terms, the

irreducible or random error, the bias (or squared bias) and the variance. The irreducible error is independent of the learning algorithm and places an upper bound on the performance of any regression technique. The bias term is a measure of how closely the learning algorithm's mean prediction over all training sets of fixed size, is near to the target. The variance is a measure of how the learning algorithms predictions for a given target, vary around the mean prediction. The purpose of an ensemble is to try to reduce bias and/or variance in the error. For a linear combination of 1,..,*N* base models where each i^{th} base model's contribution to the ensemble prediction is weighted by a coefficient α_i and $\sum_{i=1..N} \alpha_i = 1$, Krogh & Vedelsby (1995) showed that the generalization error of the ensemble trained on a single data set can also be decomposed into two terms. The first term consists of the weighted error of the individual ensemble members (their *weighted error or average accuracy*) and the second term represents the variability of the ensemble members predictions referred to as the *ambiguity* (or diversity) of the ensemble. They demonstrated that as the ambiguity increases and the first term remains the same, the error of the ensemble decreases. Brown et al. (2005) extended this analysis by looking at the bias-variance decomposition of a similarly composed ensemble and determined that the expected generalization error of the ensemble (if each model was equally weighted) can be decomposed into the expected average individual error and expected ambiguity and showed that these terms are not completely independent. They showed that increasing ambiguity will lead to a reduction in the error variance of the ensemble, but it can also lead to an increase in the level of averaged error in the ensemble members. So, in effect, there is a trade off between the ambiguity/diversity of an ensemble and the accuracy of its members.

MAIN FOCUS

In this section we consider in detail ensemble generation and integration methods.

Ensemble Generation

Ensembles can be generated to increase the level of diversity in homogeneous base models using the following methods:

Vary the learning parameters: If the learning algorithm has learning parameters, set each base model to have different parameter values e.g. in the area of neural networks one can set each base model to have different initial random weights or a different topology (Sharkey, 1999) or in the case of regression trees, each model can be built using a different splitting criteria or pruning strategy. In the technique of random forests (Breiman, 2001), regression trees are grown using a random selection of features for each splitting choice or in addition, randomizing the splitting criteria itself (Geurts et al., 2006).

Vary the data employed: Each base model is built using samples of the data. Resampling methods include cross-validation, boot-strapping, sampling with replacement (employed in Bagging (Breiman, 1996a), and adaptive sampling (employed in Boosting methods). If there are sufficient data, sampling can be replaced by using disjoint training sets for each base model.

Vary the features employed: In this approach, each model is built by a training algorithm, with a variable sub-set of the features in the data. This method was given prominence in the area of classification by Ho (Ho 1998a; Ho 1998b) and consists of building each model using training data consisting of input features in a *r*-dimensional random subspace subset from the original *p*-dimensional feature space. Tsymbal et al. (2003) proposed a variant of this approach to allow variable length feature sub-sets.

Randomised Outputs: In this approach, rather than present different regressors with different samples of input data, each regressor is presented with the same training data, but with output values for each instance perturbed by a randomization process (Breiman, 2000).

Ensemble Integration

The integration of ensembles works by either combining the base models outputs in some fashion or using selection methods to choose the "best" base model. Specific combination/selection methods which learns a meta-model as an integration technique, are described as meta-methods.

Much of the early research involving ensemble learning for regression focused on the combination of ANNs (Sharkey, 1999). However many of these methods can be applied directly to any ensemble of regression models, regardless of the base algorithm