Neurocomputing 000 (2018) 1–14

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Diversity and degrees of freedom in regression ensembles

Henry WJ Reeve*, Gavin Brown

University of Manchester - School of Computer Science Kilburn Building, University of Manchester, Oxford Rd, Manchester M13 9PL, United Kingdom

ARTICLE INFO

Article history:
Received 10 July 2017
Revised 27 November 2017
Accepted 28 December 2017
Available online xxx

Keywords:
Degrees of freedom
Negative Correlation Learning
Tikhonov regularisation
Ensembles
Stein's unbiased risk estimate
Deep neural networks

ABSTRACT

Ensemble methods are a cornerstone of modern machine learning. The performance of an ensemble depends crucially upon the level of diversity between its constituent learners. This paper establishes a connection between diversity and degrees of freedom (i.e. the capacity of the model), showing that diversity may be viewed as a form of *inverse regularisation*. This is achieved by focusing on a previously published algorithm *Negative Correlation Learning* (NCL), in which model diversity is explicitly encouraged through a diversity penalty term in the loss function. We provide an exact formula for the effective degrees of freedom in an NCL ensemble with fixed basis functions, showing that it is a continuous, convex and monotonically increasing function of the diversity parameter. We demonstrate a connection to Tikhonov regularisation and show that, with an appropriately chosen diversity parameter, an NCL ensemble can always outperform the unregularised ensemble in the presence of noise. We demonstrate the practical utility of our approach by deriving a method to efficiently tune the diversity parameter. Finally, we use a Monte-Carlo estimator to extend the connection between diversity and degrees of freedom to ensembles of deep neural networks.

1. Introduction

Ensemble methods are a cornerstone of modern machine learning. Numerous applications have shown that by combining a multiplicity of models we are able to train powerful estimators from large data sets in a tractable way. Successful ensemble performance emanates from a fruitful trade-off between the individual accuracy of the models and their diversity [10]. Typically diversity is introduced implicitly, by sub-sampling the data or varying the architecture of the models. In this paper we consider Negative Correlation Learning (NCL) [32], a powerful approach to learning ensembles of neural networks, in which diversity is encouraged explicitly by appending a diversity penalty term to the loss function. In the context of the recent breakthroughs in deep neural networks, ensembles of neural networks are likely to play an increasingly prominent role in machine learning applications. Thus, it is crucial that we obtain a deeper understanding of the dynamics of ensemble methods well suited to neural networks such as NCL. The statistical properties of NCL have already been studied in some detail [10,11,32]. Nonetheless, important questions remain surrounding the *diversity parameter*, the central hyperparameter in NCL which controls the level of emphasis placed upon the diversity penalty term. We shall address the following:

- How does the complexity of the ensemble estimator vary as a function of the diversity parameter?
- How can we efficiently optimise the diversity parameter on large data sets?
- Is the optimal value of the diversity parameter always strictly less than one?

The core of our investigation lies in a degrees of freedom analysis of NCL ensembles. Our contributions are as follows:

- We derive a formula for the degrees of freedom under the assumption of fixed basis functions (Section 3).
- We show analytically that the degrees of freedom is monotonically increasing as a function of the diversity parameter (Section 3).
- We present the surprising result that, in the presence of noise, the optimal value of the diversity parameter is always strictly less than one (Section 4).
- We develop an intriguing connection between NCL and Tikhonov regularisation (Section 5).
- We present an empirical verification of the theoretical results (Section 6).
- We give a fast and effective procedure for tuning the diversity parameter based upon the degrees of freedom (Section 7).
- We investigate ensembles of deep neural networks, demonstrating empirically that the degrees of freedom also behave monotonically with respect to the diversity parameter in this setting (Section 8).

* Corresponding author.
E-mail address: henry.reeve@manchester.ac.uk (H.W. Reeve).

https://doi.org/10.1016/j.neucom.2017.12.066
0925-2312/

The present paper extends a previously published conference paper [40]. The previous conference paper introduces the analytic formula for the degrees of freedom and demonstrates a computationally efficient approach to tuning the diversity parameter based on the formula. In the present paper we have extended this work. Firstly, we present additional technical results: A connection between NCL and Tikhonov regularisation and a result implying that the diversity parameter should never be set to precisely one in the presence of noise. Secondly, we used a Monte-Carlo estimator to conduct a detailed empirical investigation into the relationship between the diversity parameter and degrees of freedom in ensembles of deep neural networks.

We shall begin by introducing the background on ensemble learning and degrees of freedom in Section 2.

2. Background

In this section we shall introduce the relevant background on Negative Correlation Learning (NCL) and degrees of freedom. We begin by setting the scene. Throughout this paper we consider the regression problem: We are given a data set $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$ with $(x_n, y_n) \in \mathcal{X} \times \mathbb{R}$. We shall assume that there is an underlying function $\mu : \mathcal{X} \to \mathbb{R}$ such that for each n, $y_n = \mu(x_n) + \epsilon_n$, where $(\epsilon_n)_{n=1}^N$ is a mean zero, independent and identically distributed random process. Our goal is to use the data $\mathcal{D}$ to provide an estimator $\hat{\mu} : \mathcal{X} \to \mathbb{R}$ of the underlying function μ.

2.1. Ensembles, diversity and the ambiguity decomposition

Ensemble methods aggregate the predictions of a multiplicity of constituent models in order to provide a more powerful model with lower generalisation error. In order for an ensemble to outperform a single model it is essential for its constituent models to be *diverse* [8]. In the classification setting, there is no straightforward relationship between the performance of an ensemble and its diversity [17]; the ensemble error can even exceed the average error of its constituent learners. In the regression setting, however, the squared error of the ensemble may be decomposed into the average squared error of its constituents minus the variance over the ensemble's predictions. To be precise, suppose we have an ensemble $\mathcal{F} = \{f_m\}_{m=1}^M$ consisting of M functions $f_m : \mathcal{X} \to \mathbb{R}$. We let $F := (1/M) \cdot \sum_{m=1}^M f_m$ denote the ensemble function. For each $(x, y) \in \mathcal{X} \times \mathbb{R}$ we have,

$$\overbrace{(F(x)-y)^2}^{\text{ensemble error}} = \overbrace{\frac{1}{M}\sum_{m=1}^{M}(f_m(x)-y)^2}^{\text{average error}} - \overbrace{\frac{1}{M}\sum_{m=1}^{M}(f_m(x)-F(x))^2}^{\text{diversity}}. \quad (1)$$

This relationship is known as the *ambiguity decomposition*. It was observed by Krogh and Vedelsby who highlighted its importance for ensemble learning [28]. We refer to the variance over the ensemble's outputs as the *diversity*. The ambiguity decomposition shows that the square error of the ensemble never exceeds the average error of its constituent learners, and the extent to which the ensemble outperforms its constituents is quantified by its diversity.

Hence, ensemble methods succeed by attaining a high degree of diversity without sacrificing too much individual accuracy. Typically ensemble methods encourage diversity implicitly by modifying the training data or model-structure for the constituent models. For example, Ada-boost encourages diversity by increasing the weight of examples mis-classified by previous models [15], whereas random forests encourage diversity by training different trees with different bootstrap samples of the data and splitting branches along different subsets of features. However, the ambiguity decomposition (1) motivates a more direct approach: *Negative Correlation Learning*.

2.2. Negative Correlation Learning

The Negative Correlation Learning method, introduced by Liu and Yao [32], encourages diversity explicitly by incorporating a diversity penalty term into the cost function. Suppose we have an ensemble $\mathcal{F} = \{f_m\}_{m=1}^M$ with each function $f_m : \mathcal{X} \to \mathbb{R}$ chosen from a parameterisable family of neural networks $\mathcal{H}_m$, parameterised by θ_m. Our ensemble estimator $\hat{\mu}$ is given by the average $F = (1/M) \cdot \sum_{m=1}^M f_m$. The NCL rule, introduced by Liu and Yao [32] proceeds as follows. First each parameter vector θ_m is randomly initialised. Then, for each training example $(x_n, y_n) \in \mathcal{D}$, we update each θ_m in parallel according to

$$\theta_m \leftarrow \theta_m - \alpha \cdot \frac{\partial f_m}{\partial \theta_m} \cdot \left(\overbrace{(f_m(x_n) - y)}^{\text{accuracy}} - \lambda \cdot \overbrace{(f_m(x_n) - F(x_n))}^{\text{diversity}} \right),$$

where α is a learning rate. Thus, each update consists of two components: The first pushes the output of the model in the direction of the target, making the model more accurate over the training data. The second pushes the individual model output away from average output, encouraging diversity. The NCL rule is equivalent to stochastic gradient descent with respect to the following loss function (with a scaled learning rate),

$$L_\lambda(\mathcal{F}, x, y) := \overbrace{\frac{1}{M}\sum_{m=1}^{M}(f_m(x)-y)^2}^{\text{accuracy}} - \lambda \cdot \overbrace{\frac{1}{M}\sum_{m=1}^{M}(f_m(x)-F(x))^2}^{\text{diversity}}. \quad (2)$$

We shall refer to L_λ as the NCL loss.

The study of NCL is important for several reasons. Firstly the NCL method has been shown to perform well on a wide variety of regression problems, in some cases significantly outperforming other ensemble methods such as boosting and bagging [10,32]. Secondly, NCL holds a privileged place amongst ensemble methods due to its explicit emphasis upon diversity. Thirdly, the past decade has seen phenomenal progress in deep learning with artificial neural networks surpassing human performance on certain tasks [4,18,27,30,41]. NCL is specifically designed for generating ensembles of neural networks. Hence, there is a great potential for future applications of NCL to deep neural networks.

The key focus of this paper will be understanding the behaviour of the ensemble as a function of the diversity parameter λ, which explicitly manages a trade-off between the two competing objectives of accuracy and diversity. An important observation of Brown et al. is that the NCL loss may be rewritten in the following way [9,10]:

$$L_\lambda(\mathcal{F}, x, y) := (1-\lambda) \cdot \overbrace{\frac{1}{M}\sum_{m=1}^{M}(f_m(x)-y)^2}^{\text{individual accuracy}} + \lambda \cdot \overbrace{(F(x)-y)^2}^{\text{combined accuracy}}. \quad (3)$$

Hence, when $\lambda = 0$ each function f_m is trained individually and when $\lambda = 1$, L_λ is the squared error for the average F. Hence, NCL scales smoothly between training each of the functions f_m individually and training as a single combined estimator F. Brown et al. conducted a detailed analysis of NCL, relating the behaviour of the ensemble to the bias-variance-covariance decomposition [10]. In addition, Brown et al. gave an upper bound on the diversity parameter λ, showing that for $\lambda > M/(M-1) > 1$ the Hessian matrix of the weights is non-positive semi-definite. It was subsequently shown that for any $\lambda > 1$, minimising L_λ causes the weights to diverge [39, Theorem 3]. Thus, we should restrict the diversity parameter λ to the region $\lambda \in [0, 1]$. Nonetheless, many open questions remain.

Please cite this article as: H.W. Reeve, G. Brown, Diversity and degrees of freedom in regression ensembles, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2017.12.066