Information Sciences 417 (2017) 55–71

Contents lists available at ScienceDirect

Information Sciences

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

Stochastic configuration networks ensemble with heterogeneous features for large-scale data analytics

Dianhui Wang*, Caihao Cui

Department of Computer Science and Information Technology, La Trobe University, Melbourne, VIC 3086, Australia

A R T I C L E   I N F O

Article history:
Received 8 December 2016
Revised 2 July 2017
Accepted 3 July 2017
Available online 4 July 2017

Keywords:
Stochastic configuration networks
Large-scale data analytics
Heterogeneous features
Ensemble learning
Negative correlation learning

A B S T R A C T

This paper presents a fast decorrelated neuro-ensemble with heterogeneous features for large-scale data analytics, where stochastic configuration networks (SCNs) are employed as base learner models and the well-known negative correlation learning (NCL) strategy is adopted to evaluate the output weights. By feeding a large number of samples into the SCN base models, we obtain a huge sized linear equation system which is difficult to be solved by means of computing a pseudo-inverse used in the least squares method. Based on the group of heterogeneous features, the block Jacobi and Gauss–Seidel methods are employed to iteratively evaluate the output weights, and a convergence analysis is given with a demonstration on the uniqueness of these iterative solutions. Experiments with comparisons on two large-scale datasets are carried out, and the system robustness with respect to the regularizing factor used in NCL is given. Results indicate that the proposed ensemble learning techniques have good potential for resolving large-scale data modelling problems.

1. Introduction

Machine learning has received considerable attention over the past years due to its significant role for data analytics [13]. Under big data setting with decentralized information structure, advanced machine learning algorithms with robust and parallel implementations are needed along with the growth of data [24,25]. Various ensemble learning frameworks, aiming to improve the generalization performance of a learning system, have been developed over the last two decades, and many interesting ideas and theoretical works, including bagging, boosting, AdaBoost and random forests can be found in [1–7,9,12,15,16,21,23,26]. Generally speaking, learning-based ensembles share some common nature in system design, such as data sampling and the output integration. The basis of ensemble learning theory lies in a rational sampling implementation for building each base learner model, which may provide a sound predictability though learning a subset of the whole data set.

For neural network ensembles [5,12,16,23], the base models are trained by the error back-propagation (BP) algorithm and the regularizing factor used in the negative correlated cost function can be determined by the cross-validation method. Unfortunately, BP algorithm suffers from the sensitive setting of the learning rate, local minima and very slow convergence. Therefore, it is challenging to apply the existing ensemble methods for large-scale data sets. To overcome this problem, we employed random vector functional-link (RVFL) networks [11,19] to develop a fast decorrelated neuro-ensemble (termed

* Corresponding author.
E-mail address: dh.wang@latrobe.edu.au (D. Wang).

http://dx.doi.org/10.1016/j.ins.2017.07.003

DNNE) in [1]. From our experience, DNNE can perform well on smaller data sets [1,15]. However, it is quite limited for dealing with large scale data because of its high computational complexity, the scalability of numerical algorithms for the least squares solution, and hardware constraint (here mainly referring to the PC memory). Recall that physical data may come from different types of sensors, localized information source or potential features extracted from multiple runs of some certain feature selection algorithms [6,10,17,18,20,22]. Thus, for large-scale data analytics, it is useful and significant to develop a generalized neuro-ensemble framework with heterogeneous features.

This paper is built on our previous work reported in [1], which is a specific implementation of the well-known NCL learning scheme using RVFL networks with a default scope setting of the random weights and biases. From theoretical statements on the universal approximation property in [11] and our empirical results on RVFL networks in [14], the default scope setting (i.e., [−1, 1]) for the random weights and biases cannot ensure the modelling performance at all. Therefore, readers should be aware of this pitfall and must be careful in making use of our code.[1] Limits of DNNE mainly come from the following aspects: (i) the system inputs are centralized or combined with different types of features; and (ii) the analysed method of computing the output weights becomes infeasible for large-scale data sets, which is related to the nature of the base learner model (i.e., the number of nodes at the hidden layer must be sufficiently large to achieve sound performance). To relax these constraints and emphasize on the fast building of neuro-ensembles with heterogeneous features, we generalize the classical NCL-based ensemble framework into a more general form, where a set of input features are feed into the SCN base models separately. This work also provides a feasible solution by using two iterative methods for evaluating the output weights of the SCN ensemble (SCNE). In addition, some analyses and discussions on the convergence of these iterative schemes are given through a demonstration on the correlations among the iterative solutions and the pseudo-inverse solution.

The remainder of the paper is organized as follows: Section 2 provides some technical supports, including the basics of the SCN model, a generalized version of the ensemble generalization error and the negative correlation learning scheme. Section 3 describes the proposed SCNE with heterogeneous features, details two iterative learning algorithms and discusses their convergence. Section 4 reports some experimental results on two large-scale data sets, including a robustness analysis on the system performance with respect to the regularizing factor used in NCL. Section 5 concludes this paper with some remarks on further studies.

2. Technical supports

This section briefly reviews the stochastic configuration networks, extends the ensemble generalization error with heterogeneous features, followed by the negative correlation learning scheme for building ensemble models.

2.1. Revisit of stochastic configuration networks

SCNs are a class of randomized learner models which are recently developed in [28]. The unique characteristics of the SCN model, different from the classical randomized learner model (i.e., RVFL networks), is the way of generating the random input weights and biases. In contrast to RVFL networks, SCNs are built incrementally according to a supervisory mechanism, which constrains the random input weights and biases to take values in a data-dependent territory, namely stochastic configuration support (SCS). This constructive approach for building SCNs guarantees the universal approximation property of the resulting SCN model for a given nonlinear map. For the sake of completeness, we revisit the main theoretical result in Theorem 1 below.

Given a target function $f: \mathcal{R}^d \to \mathcal{R}^m$. Suppose that an SCN model has already been built with $L-1$ hidden nodes, i.e., $f_{L-1} = \sum_{l=1}^{L-1} \beta_l \phi_l(\boldsymbol{w}_l^T \boldsymbol{x} + b_l)$ $(L = 1, 2, \ldots;\ f_0 = 0)$, where $\beta_l = [\beta_{l,1}, \beta_{l,2}, \ldots, \beta_{l,m}]^T$, and $\phi_l(\boldsymbol{w}_l^T \boldsymbol{x} + b_l)$ is an activation function of the lth hidden node with random input weights $\boldsymbol{w}_l$ and bias b_l. Denoted the residual error by $e^*_{L-1} = f - f_{L-1} = [e^*_{L-1,1}, \ldots, e^*_{L-1,m}]$, where $[\beta_1^*, \beta_2^*, \ldots, \beta_{L-1}^*] = \arg\min_{\beta} \|f - \sum_{l=1}^{L-1} \beta_l \phi_l\|$.

Let $\Gamma = \{\phi_1, \phi_2, \phi_3, \ldots\}$ be a set of real-valued functions, and span(Γ) denote a function space spanned by Γ; $L_2(D)$ denote the space of all Lebesgue measurable functions $f = [f_1, f_2, \ldots, f_m]: \mathcal{R}^d \to \mathcal{R}^m$ defined on $D \subset \mathcal{R}^d$, with the L_2 norm defined as

$$\|f\| = \left(\sum_{q=1}^{m} \int_D |f_q(x)|^2 dx \right)^{1/2} < \infty. \tag{1}$$

The inner product of $\theta = [\theta_1, \theta_2, \ldots, \theta_m]: \mathcal{R}^d \to \mathcal{R}^m$ and f is defined as

$$\langle f, \theta \rangle = \sum_{q=1}^{m} \langle f_q, \theta_q \rangle = \sum_{q=1}^{m} \int_D f_q(x) \theta_q(x) dx. \tag{2}$$

Theorem 1 (Wang and Li [28]). *Suppose that span(Γ) is dense in L_2 space and for any $\phi \in \Gamma$, $0 < \|\phi\| < b_\phi$ for some $b_\phi \in \mathcal{R}^+$. Given $0 < r < 1$ and a nonnegative real number sequence $\{\mu_L\}$ with $\lim_{L\to+\infty} \mu_L = 0$ subjected to $\mu_L \le (1-r)$. For $L = 1, 2, \ldots,$*

[1] http://homepage.cs.latrobe.edu.au/dwang/html/DNNEweb/index.html.