Neurocomputing 73 (2010) 2799–2808

Contents lists available at ScienceDirect

Neurocomputing

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

An efficient multi-objective learning algorithm for RBF neural network

Illya Kokshenev *, Antonio Padua Braga

Universidade Federal de Minas Gerais, Depto. Engenharia Eletrônica Av. Antônio Carlos, 6.627 - Campus UFMG Pampulha 30, 161-970 Belo Horizonte, MG, Brazil

ARTICLE INFO

Available online 22 August 2010

Keywords:
Multi-objective learning
Radial-basis functions
Pareto-optimality
Model selection
Regularization

ABSTRACT

Most of modern multi-objective machine learning methods are based on evolutionary optimization algorithms. They are known to be global convergent, however, usually deliver nondeterministic results. In this work we propose the deterministic global solution to a multi-objective problem of supervised learning with the methodology of nonlinear programming. As the result, the proposed multi-objective algorithm performs a global search of Pareto-optimal hypotheses in the space of RBF networks, determining their weights and basis functions. In combination with the Akaike and Bayesian information criteria, the algorithm demonstrates a high generalization efficiency on several synthetic and real-world benchmark problems.

1. Introduction

Many tasks of intelligent data analysis are covered by the field of machine learning. As known, solutions to common problems of machine learning, such as pattern recognition, regression, and categorization (clustering) always result into trade-offs among several concurrent objectives of learning. For instance in supervised learning, the trade-off between the empirical risk (training error) and capacity of a hypotheses class (model complexity) is depicted by the paradigms of Statistical Learning Theory (SLT) [1] and the bias-variance dilemma [2], playing essential role in the performance of a learning machine. Namely, the principle of structural risk minimization (SRM) [3] states that the error and complexity must be minimized maintaining a certain balance in order to achieve a solution to the learning problem, characterized by good generalization properties.

The principle of SRM is usually implemented by means of error minimization while controlling the complexity of the model. Such approach is employed in many learning machines, such as neural networks with weight decay or pruning, regularization networks, and support vector machines (SVM). They minimize both error and complexity as a single loss function, whereas the point of balance is pre-determined by one or several hyperparameters (e.g., regularization and kernel parameters). Each choice of hyperparameters provides only a particular solution (learning hypothesis), which is not necessary efficient since not all choices of hyperparameters represent the trade-off between the error and complexity. In contrast, the principle of Pareto-optimality permits one to express the complete set of efficient solutions through the multi-criteria formulation of the learning problem. This approach led to a development of the multi-objective machine learning (MOML) [4].

A direct application of the Pareto-optimality principle to a general set of hypotheses usually results in non-convex problems, whose global solutions are required. Due to NP-complexity of such problems and difficulties of finding their solutions analytically, the arsenal of MOML methods went to the field of rapidly developing evolutionary multi-objective optimization (EMO) [5,6], as witnessed by the recent review on the subject [7]. In particular, most MOML algorithms (e.g., [8–12]) emerge from the genetic population-based approach. As an alternative, applications of nonlinear programming methods are demonstrated in [13–15], where the MOML problem of finding Pareto-optimal hypotheses in the domain of multilayer perceptrons (MLP) is approached with the so-called MOBJ algorithms.

The MOBJ algorithms are deterministic. However, they rely on the locally convergent optimization directly applied to generally non-convex problems, suffering from the problem of local minima. Hence, the Pareto-optimality is not guaranteed. On the other hand, the EMO algorithms are based on heuristics, providing the nondeterministic approximations of Pareto sets with populations of nondominated elements, which are unable to reach Pareto-optimality within a guaranteed time.

Despite of high capabilities of EMO, certain multi-objective problems can be efficiently solved in a deterministic way, taking advantages of nonlinear programming. In particular, the earlier proposed in [16] idea of decomposition of the multi-objective problem into a set of convex subproblems led to a development of the MOBJ algorithm for finding Pareto-optimal solutions within a small class of hypotheses of RBF networks. Such an approach allows to approximate Pareto sets arbitrary well with the numbers of exact solutions of convex subproblems.

In this work, we provide a deeper study of the previous results [16] and extend their application to larger classes of hypotheses.

* Corresponding author.
E-mail addresses: illya.kokshenev@gmail.com (I. Kokshenev), apbraga@cpdee.ufmg.br (A. Padua Braga).

0925-2312/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2010.06.022

Specifically, we show the possibility of finding Pareto-optimal hypothesis within the class of RBF networks of arbitrary structures. The proposed MOBJ algorithm determines the weights, widths, and centers of the basis functions as well as their quantity. Also, a special attention is payed to the problem of selection of the final solution (model selection) from the wide spectrum of Pareto-optimal hypotheses.

2. Multi-objective view on supervised learning

Let Ω be the set of learning hypotheses and $\phi : \Omega \to \mathbb{R}^r$, $r \in \mathbb{N}$ be the vector-function of learning objectives. Without loss of generality, we assume that all $r \geq 2$ components of ϕ are aimed for minimization under Ω. When there exists such a hypothesis $f \in \Omega$ that simultaneously turns all components of ϕ into their global extrema, the solution to the minimization problem is, obviously, f. Otherwise, the solution to the multi-objective problem is the set

$$\mathcal{P}(\Omega) := \{f \in \Omega | \forall f' \neq f \in \Omega (f \prec f')\} \tag{1}$$

of nondominated hypothesis, also known as Pareto set. Here, for two hypothesis $f \in \Omega$ and $f' \in \Omega$ we denote $f \prec f'$ with the meaning "f strictly dominates f'". In our minimization setting, $f \prec f'$ is true *iff* $\phi(f') \neq \phi(f)$ and all components of the difference vector $\phi(f') - \phi(f)$ are non-negative. In other words, $f \prec f'$ is true when the hypothesis f is not worse than f', but also is better with respect to at least one of the objectives. Hence, the nondominated elements represent a set of solutions which cannot be improved any further, thus, they are optimal, i.e., Pareto-optimal. The Pareto set can be viewed as the lower bound of Ω under the strict partial order relation $\prec$, whereas its geometry can be studied in $\mathbb{R}^r$ from its image under ϕ, denoted as $\rho(\Omega) := \phi(\mathcal{P}(\Omega))$, also known as Pareto front.

In particular, we consider the case when Ω is the set of input–output mapping functions, corresponding to a certain class of neural networks. When $\phi_e : \Omega \to \mathbb{R}$ and $\phi_c : \Omega \to \mathbb{R}$ are the empirical risk and model complexity functionals, respectively, the bi-criteria minimization problem

$$\min_{f \in \Omega} \phi(f) = (\phi_e(f), \phi_c(f))^T \tag{2}$$

corresponds to the supervised learning in its multi-objective formulation, referred to as MOBJ [13]. When Ω is an uncountable set, the Pareto front $\rho(\Omega)$ is a non-increasing curve in $\mathbb{R}^2$ and for some arbitrary ϕ may contain non-convex intervals and discontinuities, as illustrated in Fig. 1.

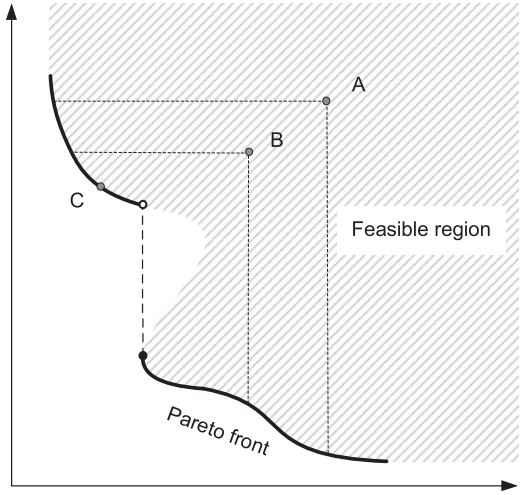

Fig. 1. Illustration of the Pareto optimality principle: the hypotheses A, B, and Pareto-optimal C are related as $C \prec B \prec A$.

The Pareto-set $\mathcal{P}(\Omega)$ usually contain infinite number of elements, equivalently efficient with respect to ϕ. Thus, it is required to make a decision towards the final hypothesis from $\mathcal{P}(\Omega)$, via application of a certain *posteriori* model selection criterion.

3. Approximations of the Pareto set

For generally non-convex objective functions, finding all Pareto-optimal hypotheses requires a global optimization, addressing the MOML to a class of NP-complete problems. Thus, approximate solutions are common in practice. In the evolutionary approach to MOML with genetic algorithms (GA) (e.g., [8,9]), the Pareto set $\mathcal{P}(\Omega)$ is approximated by a finite population of hypotheses which are getting closer to $\mathcal{P}(\Omega)$ after each evolution step. However, the elements of $\mathcal{P}(\Omega)$ can be analytically expressed as solutions of the single-objective optimization problems by means of the so-called scalarization techniques. For instance, the well-known ε−constraint [17] method determines the Pareto-optimal hypothesis of the MOBJ problem (2) as a solution of the constrained error minimization problem

$$\begin{aligned} &\min_{f \in \Omega} \quad \phi_e(f) \\ &\text{s.t.} \quad \phi_c(f) \leq \varepsilon_i. \end{aligned} \tag{3}$$

The set of solutions of (3), corresponding to a finite sequence of restriction parameters $(\varepsilon_i)_i$, is the subset of $\mathcal{P}(\Omega)$, and, thus, is its finite-set approximation $\tilde{\mathcal{P}}(\Omega) \subseteq \mathcal{P}(\Omega)$.

Another traditional scalarization method is the weighted-sum. Namely, when ϕ_e and ϕ_c are strictly convex on Ω, the minimization of their convex combination

$$\min_{f \in \Omega} \phi_e(f) + \lambda_i \phi_c(f) \tag{4}$$

is equivalent to (3) and draws Pareto-optimal elements from (2) (see e.g., [18, Chapter 3] and [19]).

Noteworthy, the commonly known learning schemes can be recognized from both (3) and (4). When ϕ_c is the measure of learning capacity of the model associated with f, the ε−constraint (3) solutions for the sequence of parameters $0 < \varepsilon_1 < \varepsilon_2 < \ldots < \infty$ minimize empirical risk ϕ_e within the structure $\emptyset \subset \Omega_1 \subset \Omega_2 \subset \ldots \subset \Omega$ of the nested subsets $\Omega_i = \{f \in \Omega | \phi_c(f) < \varepsilon_i\}$, explicitly implementing the principle of SRM. On the other hand, when (4) is minimized with $\phi_c(f)$, being a certain smoothness measure of f, one recovers a certain form of the regularization [20,21] in Ω. However, the latter requires a strict convexity of (4) for holding its equivalence to (2).

Usually, one is interested in Ω to be a class of universal approximators, e.g., neural networks of all possible topologies up to a certain size. Obviously, in this case ϕ_e contains multiple local minima and, consequently, is non-convex on Ω. Hence, due to the convexity limitations of the weighted-sum the Pareto set $\mathcal{P}(\Omega)$ cannot be entirely approximated with (4), whereas application of (3) requires globally convergent optimization procedures. Instead, following the earlier ideas from [16], we propose the decomposition of the problem domain by the union

$$\Omega = \bigcup_i \Omega_i.$$

Given that $\mathcal{P}(\mathcal{P}(\Omega)) = \mathcal{P}(\Omega)$, one can infer that $\mathcal{P}(A \cup B) = \mathcal{P}(\mathcal{P}(A) \cup \mathcal{P}(B))$ and thereby find the Pareto set from the relation

$$\mathcal{P}(\Omega) = \mathcal{P}\left(\bigcup_i \mathcal{P}(\Omega_i)\right). \tag{5}$$

When the subsets Ω_i are such that ϕ_e and ϕ_c are strictly convex under Ω_i, the elements of $\mathcal{P}(\Omega_i)$ in (5) can be efficiently found by