Knowledge-Based Systems 95 (2016) 86–98

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Learning and approximation capabilities of orthogonal super greedy algorithm☆

Jian Fang[a], Shaobo Lin[b,*], Zongben Xu[a]

[a] Institute for Information and System Sciences, School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China
[b] College of Mathematics and Information Science, Wenzhou University, Wenzhou 325035, China

ARTICLE INFO

Article history:
Received 9 April 2015
Revised 16 November 2015
Accepted 20 December 2015
Available online 31 December 2015

Keywords:
Supervised learning
Nonlinear approximation
Orthogonal super greedy algorithm
Orthogonal greedy algorithm

ABSTRACT

We consider the approximation capability of orthogonal super greedy algorithms (OSGA) and its applications in supervised learning. OSGA focuses on selecting more than one atoms in each iteration, which, of course, reduces the computational burden when compared with the conventional orthogonal greedy algorithm (OGA). We prove that even for function classes that are not the convex hull of the dictionary, OSGA does not degrade the approximation capability of OGA, provided the dictionary is incoherent. Based on this, we deduce tight generalization error bounds for OSGA learning. Our results show that in the realm of supervised learning, OSGA provides a possibility to further reduce the computational burden of OGA on the premise of maintaining its prominent generalization capability.

1. Introduction

A greedy algorithm is a stepwise inference process that follows the problem solving heuristic of making the locally optimal choice at each step with the hope of finding a global optimum. The use of greedy algorithms in the context of nonlinear approximation [1] is very appealing since it greatly reduces the computational burden when compared with standard model selection method using general dictionaries. This property triggers avid research activities of greedy algorithms in signal processing [7,17,31], inverse problem [13,32] and sparse approximation [12,29].

Greedy learning, or more specifically, applying greedy algorithms to tackle supervised learning problems, has been proved to possess charming generalization capability with lower computational burden than the widely used coefficient-based regularization methods [1]. From approximation to learning, greedy learning can be usually formulated as a four-stage stepwise learning strategy [35]. The first one is the "dictionary-selection" stage which aims at selecting a suitable set of candidates to build up the dictionary. The second one is the "greedy-definition" stage that sets the measurement criterion to choose new atoms (or elements) from the dictionary in each greedy step. The third one is the "iterative-rule" stage that defines the estimator based on the selected "greedy atoms" and the estimator obtained in the previous greedy step. The last one is the "stopping-criterion" stage which focuses on how to terminate the learning process.

Since greedy learning's inception in supervised learning [14], the aforementioned four stages were comprehensively studied for various purposes. For the "dictionary-selection" stage, Chen et al. [4] and Lin et al. [18] proposed that the kernel based dictionary is a good choice for greedy learning. For the "greedy-definition" stage, Xu et al. [35] pointed out that the metric of greedy-definition is not uniquely the greediest one. They provided a threshold to discriminate whether a selection is greedy and analyzed the feasibility of such a discrimination measurement. For the "iterative-rule" stage, Barron et al. [1] declared that both relaxed greedy iteration and orthogonal greedy iteration can achieve fast learning rates for greedy learning. For the "stopping-criterion" stage, Barron et al. [1] provided an l^0 complexity regularization strategy and Chen et al. [4] proposed an l^1 complexity constraint strategy. All these results showed that as a feasible learning scheme, greedy learning deserves comprehensively studying due to its stepwise learning character [14].

Although the importance of a single stage of greedy learning was revealed [1,4,18,34], the relationship between these stages and their composite effects for learning also need classifying. In the recent work [35], Xu et al. established a relationship between the "greedy-definition" and "stopping-criterion" stages and successfully reduced the computational cost of greedy learning without sacrificing the generalization capability very much. This implies that

☆ The research was supported by the National 973 Programming (2013CB329404), and the National Natural Science Foundation of China (Grant Nos. 61502342, 11401462).

* Corresponding author. Tel.: +8618267815286.
E-mail address: sblin1983@gmail.com (S. Lin).

http://dx.doi.org/10.1016/j.knosys.2015.12.011
0950-7051/

the study of these relationships may bring additional benefits of greedy learning. In this paper, we aim to study the relationship between the "dictionary-selection" and "greedy-definition" stages of orthogonal greedy algorithms (OGA). Our idea mainly stems from an interesting observation. We observe that if the selected dictionary is an orthogonal basis, then it is not necessary to define greedy learning as a stepwise strategy. Indeed, due to the orthogonal property, we can select all required atoms from the dictionary simultaneously. Conversely, if the dictionary is redundant (or linear dependent), then greedy learning must be defined as a stepwise strategy due to the redundant property. This implies that certain specific features of a dictionary can be employed to modify the greedy definition.

Therefore, if the coherence, a specific feature of a dictionary, is utilized to describe the dictionary, we can improve the performance of OGA in the direction of either reducing the computational burden or enhancing the generalization capability. In this paper, we study the learning capability of orthogonal super greedy algorithm (OSGA) which was proposed by Liu and Temlyakov [19]. OSGA selects more than one atoms from a dictionary in each iteration and hence reduces the computational burden of OGA. The aim of the present paper can be explained in two folds. The first one is to study the approximation capability of OSGA and the other is to pursue its applications in supervised learning.

For OSGA approximation, it was shown in [19] (see also [20]) that with incoherent dictionaries, OSGA can reduce the computational burden when compared with OGA. It can be found in [19, Theorem 2] that such a significant computational burden-reduction does not degrade the approximation capability if the target functions belong to the convex hull of the dictionary. However, such an assumption to the target functions is very stringent if the dimension of variable is large [1]. Our purpose is to circumvent the above problem by deducing convergence rates for functions not simply related to the convex hull of the dictionary. Interestingly, we find that, even for functions out of the convex hull of the dictionary, the approximation capability of OSGA is similar as that of OGA [1].

For OSGA learning, we prove that if the dictionary is incoherent, then OSGA learning with appropriate step-size can reduce the computational burden of OGA learning further. In particular, using the established approximation results of OSGA, we can deduce an almost same learning rate as that of OGA. This means that studying the relationship between the "dictionary-selection" and "greedy-definition" stages can build more efficient learning schemes than OGA. Both numerical simulations and real data experiments illustrate the outperformance of OSGA and therefore, verify our theoretical assertions.

The paper is organized as follows. In Section 2, we review notations and preliminary results in greedy-type algorithms that are frequently referred to throughout the paper. In Section 3, we show the main results of this paper, including a general approximation theorem for OSGA and its applications in supervised learning. In Sections 4 and 5, we present a line of simulations and real data experiments to verify our viewpoints. In Section 6, we present proofs of the main results. In the last section, we further discuss the OSGA learning and draw a simple conclusion of this paper.

2. Greedy-type algorithms

Let H be a Hilbert space endowed with norm and inner product $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$, respectively. Let $\mathcal{D}=\{g\}_{g\in\mathcal{D}}$ be a given dictionary. Define $\mathcal{L}_1=\{f: f=\sum_{g\in D}a_g g\}$. The norm of $\mathcal{L}_1$ is defined by $\|f\|_{\mathcal{L}_1}:=\inf\left\{\sum_{g\in\mathcal{D}}|a_g|: f=\sum_{g\in\mathcal{D}}a_g g\right\}$. We shall assume here and later that the elements of the dictionary are normalized according to $\|g\|=1$.

There exist several types of greedy algorithms [27]. The four most commonly used are the pure greedy, orthogonal greedy, relaxed greedy and stepwise projection algorithms, which are often denoted by their acronyms PGA, OGA, RGA and SPA, respectively. In all the above greedy algorithms, we begin by setting $f_0:=0$. The new approximation $f_k\,(k\geq 1)$ is defined based on f_{k-1} and its residual $r_{k-1}:=f-f_{k-1}$. In OGA, f_k is defined as

$$f_k=P_{V_k}f,$$

where P_{V_k} is the orthogonal projection onto $V_k=\mathrm{span}\{g_1,\ldots,g_k\}$ and g_k is defined as

$$g_k=\arg\max_{g\in\mathcal{D}}|\langle r_{k-1},g\rangle|.$$

Let

$$M=M(\mathcal{D})=\sup_{g\neq h,g,h\in\mathcal{D}}|\langle g,h\rangle|$$

be the coherence of the dictionary $\mathcal{D}$. Let $s\geq 1$ be a natural number. Initially, set $f_0^s=0$ and $r_0^s=f$, then the OSGA proposed in [19] for each $k\geq 1$ can be inductively define as the following.

(1) $g_{(k-1)s+1},\ldots,g_{ks}\in\mathcal{D}$ are chosen according to

$$\min_{i\in I_k}|\langle r_{k-1}^s,g_i\rangle|\geq\sup_{g\in\mathcal{D},g\neq g_i,i\in I_k}|\langle r_{k-1}^s,g\rangle|,$$

where $I_k=[(k-1)s+1,ks]$.

(2) Let $V_{ks}=\mathrm{span}\{g_1,\ldots,g_{ks}\}$ and define

$$f_k^s:=P_{V_{ks}}f, \tag{2.1}$$

and

$$r_k^s=f-f_k^s.$$

The following Lemma 2.1 proved in [19] shows that OSGA can achieve the optimal approximation rate of ks term nonlinear approximation [26].

Lemma 2.1. *Let $\mathcal{D}$ be a dictionary with coherence M. Then, for $s\leq(2M)^{-1}+1$, the OSGA estimator (2.1) provides an approximation of $f\in\mathcal{L}_1$ with the following error bound:*

$$\|r_k^s\|^2\leq 40.5\|f\|_{\mathcal{L}_1}(sk)^{-1},\quad k=1,2,\ldots.$$

3. Approximation and learning by OSGA

In this section, after presenting some basic conceptions of the statistical learning theory, we deduce a general approximation theorem concerning OSGA and pursue its applications in regression.

3.1. Statistical learning theory

In most of machine learning problems, data are taken from two sets: the input space $X\subseteq\mathbf{R}^d$ and the output space $Y\subseteq\mathbf{R}$. The relation between the variable $x\in X$ and the variable $y\in Y$ is not deterministic, and is described by a probability distribution ρ on $Z:=X\times Y$ that admits the decomposition

$$\rho(x,y)=\rho_X(x)\rho(y|x),$$

in which $\rho(y|x)$ denotes the conditional (given x) probability measure on Y, and ρ_X the marginal probability measure on X. Let $\mathbf{z}=(x_i,y_i)_{i=1}^n$ be a set of finite random samples of size n, $n\in\mathbf{N}$, drawn identically and independently according to ρ from Z. The set of examples $\mathbf{z}$ is called a training set. Without loss of generality, we assume that $|y_i|\leq L$ for a prescribed (and fixed) $L>0$.

The goal of regression is to derive a function $f: X\to Y$ from a training set such that $f(x)$ is an effective and reliable estimate of y when x is given. A natural measurement of the error incurred by using $f(x)$ for this purpose is the generalization error, given by

$$\mathcal{E}(f):=\int_Z(f(x)-y)^2d\rho,$$