Neural Networks 49 (2014) 59–73

Contents lists available at ScienceDirect

Neural Networks

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

Generalization ability of fractional polynomial models

Yunwen Lei[a,*], Lixin Ding[a], Yiming Ding[b]

[a] State Key Lab of Software Engineering, School of Computer, Wuhan University, Wuhan 430072, China

[b] Wuhan Institute of Physics and Mathematics, Chinese Academy of Sciences, Wuhan 430071, China

ARTICLE INFO

Article history:
Received 15 January 2013
Received in revised form 23 July 2013
Accepted 22 September 2013

Keywords:
Learning algorithm
Learning theory
Fractional polynomial
Model selection
Approximation theory

ABSTRACT

In this paper, the problem of learning the functional dependency between input and output variables from scattered data using fractional polynomial models (FPM) is investigated. The estimation error bounds are obtained by calculating the pseudo-dimension of FPM, which is shown to be equal to that of sparse polynomial models (SPM). A linear decay of the approximation error is obtained for a class of target functions which are dense in the space of continuous functions. We derive a structural risk analogous to the *Schwartz Criterion* and demonstrate theoretically that the model minimizing this structural risk can achieve a favorable balance between estimation and approximation errors. An empirical model selection comparison is also performed to justify the usage of this structural risk in selecting the optimal complexity index from the data. We show that the construction of FPM can be efficiently addressed by the variable projection method. Furthermore, our empirical study implies that FPM could attain better generalization performance when compared with SPM and cubic splines.

1. Introduction

Given some scattered noisy examples $(x_i, y_i)_{i=1}^l$, the main goal of learning is to fit a function (model) to reveal the relationship between the input and output variables. Such kind of fitting should be predictive in the sense that it can estimate the outputs well for the previously unseen data (Poggio & Smale, 2003). The generalization error, which reflects the expected risk when using the obtained model to imitate the process of generating the sample, is critically important for the predictability of the constructed model (e.g., Niyogi & Girosi, 1999; Vapnik, 2000). Consequently, analyzing the generalization error occupies a central place in learning theory.

A standard approach to addressing the generalization error is to decompose it into two parts: the estimation error and the approximation error (Niyogi & Girosi, 1996, 1999). These two kinds of errors come from two different factors which are responsible for the model's generalization ability: (1) the insufficient information about the underlying input–output variable relationship due to the limited number of examples, and (2) the insufficient representational capacity of the hypothesis space where the learning process is implemented (Niyogi & Girosi, 1996, 1999). A model with good generalization performance should attain a favorable balance between the estimation and approximation errors (Geman, Bienenstock, & Doursat, 1992; Györfi, Kohler, Krzyżak, & Walk, 2002; Hastie, Tibshirani, & Friedman, 2001).

For some specific learning methods, results on characterizing their generalization ability have been obtained. Cucker and Smale (2002) and Cucker and Zhou (2007) provided a systematic study on generalization errors when the hypothesis space is a *Reproducing Kernel Hilbert Space* (RKHS). It is well known that a Hilbert space with bounded evaluation functionals possesses a reproducing kernel (Aronszajn, 1950), thus the hypothesis space considered by Cucker and Smale and Cucker and Zhou is rather general. Niyogi and Girosi (1999) characterized the generalization ability of hypothesis spaces consisting of linear superpositions of nonlinearly parameterized functions, which include radial basis functions and multilayer perceptrons as specific instances. For some classes of artificial neural networks, Bartlett, Maiorov, and Meir (1998), Haussler (1992) and Krzyżak and Linder (1998) presented some effective upper bounds for estimation errors based on an extension of the *Probably Approximately Correct* (PAC) model, while their approximation power has been extensively studied in Barron (1993) and Niyogi and Girosi (1999).

In this paper, we study the generalization performance of fractional polynomial models (FPM) and their applications to regression problems. A function $f(x)$ is called a fractional polynomial if it can be expressed with the form:

$$f(x) = \sum_{i=1}^{n} c_i x^{t_i}, \quad c_i, t_i \in \mathbb{R}, \ x \in \mathbb{R}^+,$$

[*] Corresponding author. Tel.: +86 027 68776032; fax: +86 027 68772519.
E-mail addresses: ywlei@whu.edu.cn (Y. Lei), lxding@whu.edu.cn (L. Ding), ding@wipm.ac.cn (Y. Ding).

0893-6080/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.neunet.2013.09.009

where $\mathbb{R}^+$ denotes the set of positive numbers. Fractional polynomials generalize the traditional sparse polynomials of the form

$$f(x)=\sum_{i=1}^{n} c_i x^{d_i}, \quad c_i \in \mathbb{R},\ d_i \in \mathbb{N}^+ \cup \{0\},\ x \in \mathbb{R},$$

by allowing the exponents t_i to take real values. In comparison with sparse polynomial models (SPM), FPM are more adaptive since the exponents are allowed to change according to the sample, which implies the possible superiority of FPM when applied to the practical problems. To the authors' best knowledge, the research on the approximation power of FPM can be dated back to Müntz, who showed that (DeVore & Lorentz, 1993) for a nonnegative sequence $0 = t_0 < t_1 < t_2 < \cdots$, the necessary and sufficient condition under which the linear combination of x^{t_i}, $i = 0, 1, \ldots$ is dense on $\mathcal{C}[0, 1]$ is that $\sum_{i=1}^{\infty} 1/t_i = \infty$. Here $\mathcal{C}[0, 1]$ denotes the collection of all continuous functions defined on [0, 1]. Royston and Altman (1994) also considered a related polynomial modeling strategy, for which the exponents are restricted to a small predefined set of integer and non-integer values.

Since FPM include SPM as a specific subset, approximation errors would be smaller in our case. However, to justify the generalization ability of the constructed model, one has to take into account the model's complexity. Indeed, the learning process performed in a class with too large capacity can lead to the overfitting phenomenon, where the model fits the sample well but has little generalization performance for the upcoming data (e.g., Cherkassky & Ma, 2009; Hastie et al., 2001). In this paper, the capacity of such hypothesis space is characterized in terms of its pseudo-dimension, which is shown to be equal to that of SPM. Consequently, allowing the exponents to take values over $\mathbb{R}$ does not increase the model's complexity in principle. The linear decay of the approximation error is derived for a class of target functions that are dense in $\mathcal{C}[a, b]$, $a > 0$. Our theoretical discussion suggests a structural risk analogous to the *Schwartz Criterion* and we have justified its application in selecting the optimal model, both theoretically and empirically. We also show that the construction of FPM can be considered as a separable nonlinear least squares problem, which can be efficiently approached by the variable projection method (Golub & Pereyra, 1973). Some experimental results are also provided to illustrate the possible advantage of FPM over the more established SPM and cubic splines.

This paper is organized as follows. Section 2 provides the statement of the problem. Section 3 and Section 4 address estimation and approximation errors, respectively. A structural risk is proposed and analyzed from a theoretical perspective in Section 5. Section 6 justifies the learning ability of FPM, as well as the usage of our structural risk in model selection, by performing some empirical comparisons. Some conclusions and interesting problems for further research are presented in Section 7.

2. Statement of the problem

Suppose the examples $\boldsymbol{z} = (z_1, z_2, \ldots, z_l) = ((x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l))$ are independently drawn according to an unknown probability measure ρ on $Z = X \times Y$ with a compact metric space X and $Y = \mathbb{R}$. The goal of learning is to construct a discriminant function $f : X \to Y$ in such a way that this function should capture as much as possible the relationship underlying the data. A standard approach to addressing this problem is to employ the guideline of *Empirical Risk Minimization* (ERM) (Vapnik, 2000), for which an appropriate hypothesis space $\mathcal{H}$ and a loss function $Q(z, f)$ defined on $Z \times \mathcal{H}$ are chosen in advance, then the estimator $\hat{f}_{\mathcal{H}}$ is obtained by minimizing the empirical risk $\mathcal{E}_{\boldsymbol{z}}(f) =: \frac{1}{l}\sum_{i=1}^{l} Q(z_i, f)$ over the space $\mathcal{H}$, i.e., $\hat{f}_{\mathcal{H}} = \operatorname{argmin}_{f\in\mathcal{H}} \mathcal{E}_{\boldsymbol{z}}(f)$. To measure the quality of the obtained model, however, another quantity named the generalization error (or risk) is introduced:

$$\mathcal{E}(f) := \int_Z Q(z, f)\mathrm{d}\rho,$$

which reflects the expected error suffered from using f to do the prediction when a new example arrives. The function $f_\rho := \operatorname{argmin}_f \mathcal{E}(f)$ is referred to as the target function, where the infimum is taken over all measurable functions.

It is helpful to decompose the generalization error into two different parts:

$$\mathcal{E}(\hat{f}_{\mathcal{H}}) - \mathcal{E}(f_\rho) = \big(\mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_\rho)\big) + \Big(\mathcal{E}(\hat{f}_{\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}})\Big),$$

where $f_{\mathcal{H}} := \operatorname{argmin}_{f\in\mathcal{H}} \mathcal{E}(f)$ is the best model in $\mathcal{H}$. Here for simplicity we assume that the minimum can be attained (Cucker & Zhou, 2007). The first term $\mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_\rho)$ is called the approximation error and the other term $\mathcal{E}(\hat{f}_{\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}})$ is referred to as the estimation error (also called the sample error in Cucker & Smale, 2002; Cucker & Zhou, 2007; Zhang & Cao, 2012). The approximation error can be made as small as possible by enlarging the space $\mathcal{H}$, while on the other hand, if the space $\mathcal{H}$ is too large, then the estimation of $\hat{f}_{\mathcal{H}}$ will be harder given a limited amount of data, leading to a large estimation error (e.g., Geman et al., 1992; Hastie et al., 2001). These two contradictory errors together describe the generalization performance of the model $\hat{f}_{\mathcal{H}}$ and they can be approached in empirical process theory and approximation theory, respectively.

As the ERM model $\hat{f}_{\mathcal{H}}$ is sensitive to the size of the class $\mathcal{H}$, a more efficient learning strategy called the *Structural Risk Minimization* (SRM) method has been developed to balance the approximation and estimation errors (Vapnik, 1998, 2000). Rather than considering a fixed class $\mathcal{H}$, SRM considers a sequence of hypothesis spaces $\mathcal{H}_n$, $n \geq 1$ with increasing complexity. Within each class $\mathcal{H}_n$, a candidate prediction rule $\hat{f}_n$ is picked out by some learning process and the structural risk $\widetilde{\mathcal{E}}_{\boldsymbol{z}}(\hat{f}_n)$ is established by adding an appropriate penalty into the empirical risk to favor the simpler models. Among the candidate models $\hat{f}_1, \hat{f}_2, \ldots$, SRM chooses the one with minimal structural risk as the ultimate prediction rule. The justification of the SRM principle has attracted much attention and the readers are referred to Bartlett, Boucheron, and Lugosi (2002), Cherkassky and Mulier (2007), Vapnik (1998, 2000) and the references therein for a comprehensive treatment.

In this paper, we always assume that the space Z takes the form $Z = X \times Y = [a, b] \times [-R, R]$, where a, R are two positive numbers. Here we restrict the variable x to be positive, which is natural since $\sqrt{x}$ is meaningless when $x < 0$ and x^{-1} takes no value at the point $x = 0$. Our purpose is to illustrate the generalization performance of FPM under the SRM principle. The hypothesis spaces considered in this paper are collections of fractional polynomials of the following form:

$$\mathcal{H}_n = \left\{ f(x) = \sum_{i=1}^{n} c_i x^{t_i} \middle| x \in [a, b],\ t_i \in [T_1, T_2], \right.$$
$$\left. \sum_{i=1}^{n} |c_i| \leq M \right\}, \quad n = 1, 2, \ldots, \tag{2.1}$$

where T_1, T_2 and M are three constants defined in Theorem 9. For simplicity of later reference, we also introduce three notations: $B_0 := \max(a^{T_1}, a^{T_2}, b^{T_1}, b^{T_2})$, $B_1 := \max(|\log a|, |\log b|)$, $B = \max(MB_0, R, 1)$. The candidate prediction rule is constructed by