Neural Networks 22 (2009) 958–969

Contents lists available at ScienceDirect

Neural Networks

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

Another look at statistical learning theory and regularization

Vladimir Cherkassky[a], Yunqian Ma[b,*]

[a] Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, United States

[b] Honeywell Labs, 1985 Douglas Drive North, Golden Valley, MN 55422, United States

ARTICLE INFO

Article history:
Received 8 January 2008
Received in revised form 14 January 2009
Accepted 16 April 2009

Keywords:
Function approximation
Statistical model estimation
Model identification
Penalization
Predictive learning
Regularization
Ridge regression
Structural risk minimization
SVM regression
VC-theory

ABSTRACT

The paper reviews and highlights distinctions between function-approximation (FA) and VC theory and methodology, mainly within the setting of regression problems and a squared-error loss function, and illustrates empirically the differences between the two when data is sparse and/or input distribution is non-uniform. In FA theory, the goal is to estimate an unknown true dependency (or 'target' function) in regression problems, or posterior probability $P(y/\mathbf{x})$ in classification problems. In VC theory, the goal is to 'imitate' unknown target function, in the sense of minimization of prediction risk or good 'generalization'. That is, the result of VC learning depends on (unknown) input distribution, while that of FA does not. This distinction is important because regularization theory originally introduced under clearly stated FA setting [Tikhonov, N. (1963). On solving ill-posed problem and method of regularization. *Doklady Akademii Nauk USSR*, *153*, 501–504; Tikhonov, N., & V. Y. Arsenin (1977). *Solution of ill-posed problems*. Washington, DC: W. H. Winston], has been later used under risk-minimization or VC setting. More recently, several authors [Evgeniou, T., Pontil, M., & Poggio, T. (2000). Regularization networks and support vector machines. *Advances in Computational Mathematics*, *13*, 1–50; Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference and prediction*. Springer; Poggio, T. and Smale, S., (2003). The mathematics of learning: Dealing with data. *Notices of the AMS*, *50* (5), 537–544] applied constructive methodology based on regularization framework to learning dependencies from data (under VC-theoretical setting). However, such regularization-based learning is usually presented as a purely constructive methodology (with no clearly stated problem setting). This paper compares FA/regularization and VC/risk minimization methodologies in terms of underlying theoretical assumptions. The control of model complexity, using regularization and using the concept of margin in SVMs, is contrasted in the FA and VC formulations.

1. Introduction

The growing use of computers and database technology has resulted in explosive growth of methods for learning (or estimating dependencies) from data. The classical (parametric) statistical approach does not provide practical solutions for flexible estimation with high-dimensional data. Therefore, several diverse methodologies have emerged to address this problem. These include well-established methodologies in statistics (multivariate regression/classification, Bayesian methods), engineering (statistical pattern recognition), signal processing (wavelets) and computer science (AI and machine learning) and more recently biologically inspired developments such as artificial neural networks, fuzzy logic, and genetic algorithms. Even though all these approaches often address the same application problems, there is little agreement on the fundamental issues involved, and it leads to heuristic techniques aimed at solving specific applications.

Next we briefly review 3 major theoretical/methodological paradigms for estimating predictive models from data:

1. *Parametric estimation*. This 'classical statistical inference' approach (due to R. Fisher) assumes that the form of unknown dependency (model) is *known*, up to the value of its parameters. The goal of statistical inference is *accurate parameter estimation* using the available data. It can be easily shown, however, that the parametric setting does not yield accurate generalization with finite samples, even when the true parametric model is known (Cherkassky & Mulier, 2007).
2. *Model identification / function approximation*. The growing use of computers for data analysis in the 70's and 80's has led to the development of flexible data-driven models. Many such methods developed by 'practical' statisticians have been introduced under model identification /function approximation framework, which is essentially an extension of the classical approach (1) where the assumption of knowledge about the parametric form is relaxed. That is, the unknown target function is

[*] Corresponding author. Tel.: +1 612 951 7420; fax: +1 612 951 7438.
E-mail addresses: cherk001@umn.edu (V. Cherkassky), yunqian.ma@honeywell.com (Y. Ma).

0893-6080/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2009.04.005

specified using flexible parameterization. Classical approaches consider representations linear in parameters (i.e., polynomials, harmonic functions), whereas recent nonlinear methods include nonlinear parameterizations (such as multilayer perceptron networks, projection pursuit, multivariate adaptive regression splines etc.). However, conceptually the goal of learning remains the same as in (1), i.e., *accurate estimation of the true model.*

3. *Risk minimization approach*, aka *predictive learning*. Under this framework, the goal of learning is generalization, i.e. obtaining models providing minimal prediction risk (for future samples). This approach has been originally developed by practitioners in the field of artificial neural networks in late 1980's (with no particular theoretical justification). The theoretical framework for predictive learning known as Statistical Learning Theory or VC-theory (Vapnik, 1982), has been relatively unknown until the late 1990's, but the wide acceptance of its practical methodology called Support Vector Machines (SVM) has increased interest in this theory (Vapnik, 1998, 2000). In this paper, we use the terms VC-theory and predictive learning interchangeably, to denote a methodology for estimating models with good generalization capabilities from available data.

In summary, the objective of learning using the model identification/ function approximation approach is to estimate the true model of observed random events (presumed to exist); whereas under the predictive learning approach the goal is just to find a 'good' model (providing good generalization for future data). This distinction is critical in the context of learning with finite samples, because:

(a) One can easily show examples where a 'good' estimated model (in the sense of generalization) provides very inaccurate (poor) approximation of the true model. Moreover, even when the true parametric form of the estimated dependency is known, estimation of its parameters from finite data may lead to poor generalization (Cherkassky & Mulier, 2007).

(b) Classical statistics and function approximation rely on the notion of the true model (underlying generated data). This is clearly an additional assumption imposed on application data. In most applications, the practical objective is to find a good predictive model, and the notion of the true model (target function) is simply a theoretical construct that cannot be directly observed. In contrast, VC-setting is based on the concept of risk minimization, and does not use the notion of a true model.

Each learning paradigm is described using its own concepts and mathematical theory, i.e. classical parametric statistics, function approximation/ regularization and VC learning theory. However, the distinction between these approaches becomes blurred when they are used to motivate practical learning algorithms, for two reasons. *First*, many learning algorithms can be introduced under different frameworks. For example, least-squares minimization (for function estimation from samples) can be derived using the parametric estimation approach (via maximum likelihood arguments) under Gaussian noise assumptions. Alternatively, least-squares minimization can be introduced under the risk minimization approach. *Second,* theoretical arguments have been often used to explain and/or improve various learning heuristics (rather than to derive new learning algorithms directly from theoretical principles). For example, most neural network learning algorithms have been pioneered by engineers and psychologists (using intuitive and biological motivation), and then later 'explained' using statistical and function approximation theoretical arguments.

Currently, there is a clear agreement that the classical parametric estimation approach is not appropriate for flexible estimation with finite samples. However, there seems to be little consensus on the distinction between the FA approach and the risk minimization (VC) framework. For example, SVM methodology has been originally developed in VC-theory, and later re-introduced under the 'regularization' framework (under FA setting). Several references (Evgeniou, Pontil, & Poggio, 2000; Hastie, Tibshirani, & Friedman, 2001; Poggio & Smale, 2003) suggest that SVM is a special case of the regularization formulation. On a historical note, we note that the use of regularization techniques in the context of learning with finite samples has been known in statistics long before support vector machines. In particular, the regularization approach (predating SVM) has been widely used *only in low-dimensional settings* such as splines and various signal denoising techniques. Quoting Ripley (1996): 'Since splines are so useful in one dimension, they might appear to be the obvious methods in more. In fact, they appear to be rather restricted and little used'.

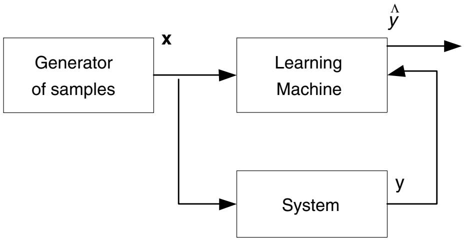

Fig. 1. Learning problem setting.

The claim about the similarity between SVM and classical regularization may suggest that there are only superficial differences between VC and FA methodologies. So, this paper is intended to clarify the differences between VC and FA approaches to learning from data. These differences are presented on 3 levels: terminology/ vocabulary, theory, and generalization performance differences. This paper is organized as follows. Section 2 presents comparisons in terms of underlying theoretical assumptions (problem setting) and mathematical concepts. Understanding these conceptual differences is important for performing meaningful empirical comparisons between regularization and a VC approach, presented in Sections 3 and 4. All empirical comparisons are performed for standard ridge regression, as a representative regularization technique. Section 3 presents comparisons between strategies for choosing regularization parameter based on the goal of FA and the goal of risk minimization (for VC approach). Section 4 shows empirical comparisons between model complexity control using the concept of margin (introduced in VC theory) and complexity control under a regularization approach. Conclusions are given in Section 5.

2. Predictive learning vs function approximation

The problem of inductive learning (or learning from examples) can be described as a generic system shown in Fig. 1 (Cherkassky & Mulier, 2007; Friedman, 1994; Vapnik, 1982). This learning system has three components:

- *Generator* of random input vectors $\mathbf{x}$, drawn independently from a fixed (but unknown) probability distribution $P(\mathbf{x})$;
- *System* (or teacher) which returns an output value y for every input vector $\mathbf{x}$ according to the fixed conditional distribution $P\,(y|\mathbf{x})$, which is also unknown;
- *Learning Machine*, or learning algorithm, which implements a set of approximating functions $f(\mathbf{x}, \omega)$, where ω is a set of parameters of an arbitrary nature.

The goal of learning is to select a function (from this set) which approximates best the System's response. This selection is based on the knowledge of a finite number (n) of samples (training data)