Guest Editorial Vapnik–Chervonenkis (VC) Learning Theory and Its Applications

***Abstract*— Statistical learning theory (aka Vapnik–Chervonenkis or VC theory) is a general mathematical framework for estimating dependencies from the empirical data. Recent interest in VC-theory has been motivated by the practical applications of a new constructive learning methodology called support vector machines (SVM's) originating from VC theory. This special issue illustrates the growing importance of VC theory for the field of predictive learning from data.**

***Index Terms*— Predictive learning, statistical estimation, support vector machines, VC theory.**

MANY neural-network applications deal with the problem of predictive learning, i.e., estimating an unknown dependency from known observations (or training samples). The same problem is also addressed in other fields such as pattern recognition and statistics. Once the dependency has been estimated, it can be used to predict future data. The problem of predictive learning is inherently difficult (ill-posed), due to the general lack of knowledge about the underlying dependency and the finiteness of available (training) data. However, with the recent growth of computers and database technology, the amount of available data has increased dramatically, and the problem of predictive learning has become increasingly important.

Currently there is no single universally accepted theoretical framework for predictive learning. Therefore, we will briefly overview several major approaches for predictive learning. We distinguish three main approaches.

- Classical (parametric) statistical estimation. Under this approach, the parametric form of the dependency is known (up to the value of its parameters). Then the training data is used to estimate the parameter values. This approach assumes strong *a priori* knowledge about the unknown dependency. Unfortunately, for many real-world problems this parametric paradigm fails to work. First, it is not practically feasible to extend the parametric approach to high-dimensional settings, since a huge number of training samples are required for accurate estimation. This is known as the curse-of-dimensionality. Second, the parametric approach is heavily based on the assumption that the underlying statistical distribution is known, whereas many real-life problems do not satisfy this assumption.
- Empirical nonlinear methods, such as artificial neural networks and flexible statistical methods were developed in 1980's to address the shortcomings of the parametric approach. These methods enable construction of nonlinear models from the available data, without making strong assumptions about the unknown dependency. However, these methods lack unified mathematical theory, and are usually motivated by biological arguments (as in the case of artificial neural networks) or by informal common-sense arguments (as flexible statistical methods, such as projection pursuit or multivariate adaptive regression splines).
- Statistical learning theory developed in the late 1960's [1] is a theory for nonparametric (distribution-free) dependency estimation with finite data. This theory is based on the theoretical analysis of the empirical risk minimization (ERM) inductive principle. The ERM principle refers to a common-sense procedure used in neural network training, where the unknown dependency is estimated via minimization of the average training error (called empirical risk) with respect to model parameters (neural-network weights). More generally, ERM is an approach for dependency estimation based on the minimization of the training error for a set of parametric functions (or approximating functions) such as an multilayer perceptron (MLP) network. The Vapnik–Cervonenkis (VC) theory derives necessary and sufficient conditions for consistency and fast rate of convergence of the ERM principle. These conditions (as most results of VC theory) are distribution-free.

Most theoretical treatments of adaptive nonlinear estimation methods (such as neural networks) follow the classical parametric framework. However, this (classical) approach appears fundamentally flawed when applied to many practical settings with finite high-dimensional data. For instance, classical notions of complexity (i.e., the number of free parameters or degrees of freedom) fail to account for successful applications of artificial neural networks with a huge number of parameters (network weights) that can successfully generalize with finite training data. In contrast, statistical learning theory uses the so-called VC dimension as a measure of complexity (capacity) of a set of approximating functions. In the simplest case of linear models the VC dimension coincides with the number of free parameters; however the VC dimension of nonlinear estimators differs from the number of parameters.

Until very recently, VC theory has not been widely applied in practice, because its original results presented in [1] and [2] are mainly of a theoretical and conceptual nature. Recent developments [3] resulted in the constructive learning

Publisher Item Identifier S 1045-9227(99)08541-0.

methodology called support vector machines (SVM's). The SVM method enables a high level of generalization with finite high-dimensional data, and hence it is very attractive for many challenging real-world applications. However, VC theory has many practical and conceptual implications beyond SVM's. We believe that VC theory can be used as a theoretical and conceptual framework for neural-network and statistical methods for learning dependencies from data. There is still much work needed to bridge the gap between theory and practical applications; however the success of SVM applications show potential advantages of developing constructive learning methods based on VC theory.

This special issue attracted 21 submitted papers from North America, Asia, and Europe. In spite of the very high quality of most submissions, only eight papers have been selected for inclusion in the special issue due to real-time constraints. The issue contains one survey paper on VC theory by Vapnik and eight papers describing various extensions and applications of SVM's. The composition of the special issue may give an unintended impression that SVM's represent the only practically useful part of VC theory. However, this is definitely not the case. An interested reader can refer to [3]–[6] for a more detailed discussion of VC theory and its relationship to statistical and neural network methods. In addition, references [5], [7], and [8] describe application of VC generalization bounds for model complexity control.

All the papers in this special issue are organized in three parts: theory and concepts, SVM algorithm implementations, and SVM applications, as described next.

I. Theory and Concepts

"An Overview of Statistical Learning Theory" by Vapnik provides a conceptual introduction to statistical learning theory. The main theme of this paper is to show how theoretical results have lead to practical learning algorithms. The article begins with a description of the learning problem which mathematically defines the problem scope. A general problem statement is provided that applies for the three commonly described learning problems: classification, regression, and density estimation. The general learning problem is described in terms of minimizing a risk functional for an unknown distribution. The difficulty is that this must be done based on the observed data (the distribution is not known). Most classical approaches for solving this problem apply the ERM principle. This principle states that we should minimize the empirical risk (i.e., training error) as a proxy for the true risk functional. The empirical risk minimization principle is very general and encompasses the classical approaches for solving learning problems, such as the least squares method in regression and the maximum likelihood method for density estimation. The article focuses on addressing the main questions of learning theory: 1) consistency and convergence properties of the empirical risk minimization principle; 2) controlling the generalization ability; and 3) constructing learning algorithms. An analytical and conceptual comparison is made between the support vector machine and sigmoid function-based multilayer perceptrons optimized via backpropagation.

"Input Space Versus Feature Space in Kernel-Based Methods" by Schölkopf *et al.* delves into the topic of feature spaces for support vector machines. It begins with a description of the geometry of feature spaces and a relevant mathematical formulation. The paper then discusses three important issues concerning kernel functions: operations in the feature space, their capacity (VC-dimension), and geometrical properties. Given a vector in the feature space, a method for determining an exact or approximate preimage in the input space is described. Preimages may be useful for denoising, compression, and feature interpretation. From preimages, the paper moves to a slightly more general problem: that of finding a reduced set expansions (i.e., an approximation using a small number of vectors in the input space) for a vector in the feature space. Reduced set expansion can be used to reduce the amount of computation required for predictions made using support vector machines.

"Moderating the Outputs of Support Vector Machine Classifiers" by Kwok extends the use of moderated outputs to the support vector machine. This is done by taking advantage of a relationship between the Bayesian evidence framework and the SVM. The moderated outputs provide an approximation to the posterior class probability, allowing the application of meaningful rejection thresholds to SVM classifiers.

II. Support Vector Machine Algorithm Implementations

"Successive Overrelaxation for Support Vector Machines" by Mangasarian and Musicant describes an alternative optimization approach for support vector classifiers with linear kernels. The optimization approach operates on one data point at a time making it applicable for huge data sets (empirical results for 10 000 000 points are shown). The algorithm converges linearly to a solution. The algorithm is compared to the sequential minimal optimization (SMO) algorithm [9] in a extensive empirical section.

"Simple and Robust Methods for Support Vector Expansions" by Mattera *et al.* describes a generalization of the support vector machine which provides a sparse solution for a linear system of equations. The paper describes two applications of sparse solutions; that of data compression and regression estimation. Experimental comparisons of the proposed SV approach and an alternative cross-correlation approach are reported.

III. Support Vector Machine Applications

"Support Vector Machines for Spam Characterization" by Drucker *et al.* provides an illustrative application example of classification using support vector machines. The paper describes some of the engineering design choices made in the development of the classifier. The support vector machine is compared to three other classification algorithms: Ripper, Rocchio, and boosting decision trees. Finally an analysis and commentary of the empirical results is presented.

In "SVM's for Histogram-Based Image Classification," Chapelle *et al.* apply support vector machines to a high-dimensional image classification task. The paper describes a

method for representing images using a histogram of the color components. Support vector machines with linear, polynomial and radial basis function kernels are compared to a k nearest neighbor technique.

"Fusion of Face and Speech Data for Person Identity Verification" by Ben-Yacoub *et al.* describes the application of support vector machines to problems in person identity authentication. The authentication is based on two modalities: face and speech. Results are presented for a number of different binary classification schemes.

REFERENCES

[1] V. Vapnik and A. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," Doklady Academii Nauk USSR, vol. 181, no. 4, 1968.
[2] V. Vapnik, *Estimation of Dependencies Based on Empirical Data*. New York: Springer-Verlag, 1982.
[3] ——, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.
[4] ——, *Statistical Learning Theory*. New York: Wiley, 1998.
[5] V. Cherkassky and F. Mulier, *Learning From Data: Concepts, Theory and Methods*. New York: Wiley, 1998.
[6] B. Schölkopf, C. Burges, and A. Smola, Eds., *Advances in Kernel Methods: Support Vector Learning*. Cambridge, MA: MIT Press, 1999.
[7] V. Cherkassky, V. X. Shao, F. Mulier, and V. Vapnik, "Model complexity control for regression using VC generalization bounds," *IEEE Trans. Neural Networks*, vol. 10, pp. 1083–1097, 1999.
[8] V. Cherkassky and X. Shao, "Model selection for wavelet-based signal estimation," in *Proc. Int. Joint Conf. Neural Networks*, Anchorage, AK, 1998.
[9] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," Microsoft Res. Tech. Rep. 98-14, Apr. 1998.

Vladimir Cherkassky, *Guest Editor*
Department of Electrical and Computer Engineering
University of Minnesota
Minneapolis, MN 55455 USA

Filip Mulier, *Guest Editor*
Net Perceptions, Inc.
Eden Prairie, MN 55344 USA