ADVANCED METHODOLOGIES for LEARNING with SPARSE DATA

Vladimir Cherkassky
Dept. ECE, University of Minnesota
Minneapolis, MN 55455
tel 612 625-9597 fax 612 625-4583
email cherk001@umn.edu
http://www.ece.umn.edu/users/cherkass/predictive_learning/

OVERVIEW: The field of Predictive Learning is concerned with estimating 'good' predictive models from available data. Such problems can be usually stated in the framework of *inductive learning*, where the goal is to estimate a good predictive model from known observations (or training data samples). In recent years, there has been a growing interest in applying learning methods to sparse high-dimensional data (i.e., in genomics, medical imaging, object recognition, etc.). In such applications, many successful approaches represent minor modifications of existing *inductive* learning methods (such as neural networks, support vector machines, discriminant analysis etc.) combined with clever preprocessing and feature extraction. At the same time, in the statistical learning community, there is a trend towards development and better understanding of new *non-standard* and *non-inductive learning settings*. Examples include (a) several powerful learning formulations developed in VC-theory: transduction, learning through contradictions, SVM+ (Vapnik, 1998, 2006); and (b) non-standard settings proposed in machine learning community, such as Multi-Task Learning (Ben-David et al, 2002), Semi-Supervised Learning (Chapelle et al, 2006) etc.. These new learning formulations are motivated by practical needs (to improve generalization for learning with sparse high-dimensional data). This tutorial will present an overview of recent non-standard learning formulations, investigate possible connections between these formulations, and discuss application examples illustrating advantages of using these approaches for sparse high-dimensional data. The presentation will be based, to a large extent, on the conceptual framework developed by Vapnik [1998, 2006].

CONTENT: This tutorial will cover three major parts. *The first part* will present VC-theoretical framework for predictive learning and standard inductive learning setting, in order to motivate alternative approaches. *Second part* presents several non-standard learning formulations such as transduction, learning through contradictions, learning with hidden information and multi-task learning. *In the third part*, we discuss practical issues and difficulties arising in application of these advanced learning techniques. These issues include model selection (parameter tuning) and interpretation of high-dimensional predictive models. Throughout this tutorial, many important points will be illustrated by empirical comparisons and related to practical applications (mainly, biomedical applications).

INTENDED AUDIENCE: Researchers and practitioners interested in understanding advanced learning methodologies, and their applications. This tutorial is also helpful for developing improved understanding of the methodological issues for learning with high-dimensional data.

References

S. Ben-David, J. Gehrke and R. Schuller, A theoretical framework for learning form a pool of disparate data sources. ACM KDD, 2002.

O. Chapelle, B. Schölkopf and A. Zien, Eds., *Semi-Supervised Learning*, MIT Press, 2006

Cherkassky, V. and F. Mulier, Learning from Data, *second edition*, Wiley, 2007

Cherkassky, V. and Y. Ma, Introduction to Predictive Learning, Springer, 2011 (to appear)

Cherkassky, V., Cai, F., and L. Liang, Predictive learning with sparse heterogeneous data, Proc IJCNN 2009

Cherkassky,V. , S. Dhar and W. Dai, "Practical Conditions for Effectiveness of the Universum Learning," IEEE Trans. on Neural Networks, (under review), 2010

Dhar, S. and V. Cherkassky, Visualization and Interpretation of SVM classifiers, Wiley Interdisciplinary Review, Data Mining and Knowledge Discovery, 2011 (to appear)

Vapnik, V., Statistical Learning Theory, Wiley, 1998

Vapnik, V., Empirical Inference Science: Afterword of 2006, Springer 2006

INSTRUCTOR: Vladimir Cherkassky is Professor of Electrical and Computer Engineering at the University of Minnesota. He received Ph.D. in Electrical Engineering from University of Texas at Austin in 1985. His current research is on methods for *predictive learning from data*, and he has co-authored a monograph *Learning From Data* published by Wiley in 1998. Prof. Cherkassky has served on the Governing Board of INNS. He has served on editorial boards of *IEEE Transactions on Neural Networks*, the *Neural Networks* Journal, the *Natural Computing* Journal and the *Neural Processing Letters*. He served on the program committee of major international conferences on Artificial Neural Networks. He was Director of NATO Advanced Study Institute (ASI) *From Statistics to Neural Networks: Theory and Pattern Recognition Applications* held in France, in 1993. He presented numerous tutorials on neural network and statistical methods for learning from data. In 2007, he became Fellow of IEEE, for 'contributions and leadership in statistical learning and neural network research'.