



HAL
open science

Editorial

Gilbert Saporta, Michel Béra

► **To cite this version:**

Gilbert Saporta, Michel Béra. Editorial. Applied Stochastic Models in Business and Industry, 2005, 21 (2), pp.95-96. 10.1002/asmb.543 . hal-03134518

HAL Id: hal-03134518

<https://cnam.hal.science/hal-03134518v1>

Submitted on 22 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Editorial

In the last 30 years, many algorithms have been developed throughout the AI/Computer Science community (artificial intelligence, neural networks, machine learning), and in Statistics (for instance: Ridge and Principal Component regression, PLS, Generalized Additive Models etc.), to solve what was called the curse of high dimensionality, where Nature embodiment of a problem, be it qualitative or quantitative, could only be described by a very large number of attributes, and observed through a relatively limited number of samples. Somehow controversially, this effort has driven the emergence of Machine Learning as a scientific community.

As years went by, powerful methodologies came to existence, and a pure mathematical background, now known as Statistical Learning Theory (SLT) was derived mainly in the 1990s. It may come as a surprise that many popular SLT developments are very close to rather old work in Multivariate Data Analysis, such as Principal Components, Multidimensional Scaling, Kernel based non-linear attributes geometries, non-parametric estimations (density, regression). Regularization techniques and control of VC-dimension are two good examples of this new way of dealing with stochastically ill-posed problems, where the dimensionality of the feature space is too high.

Today's SLT efficiency, in terms of computing complexity, allows for a huge scalability in database and problem sizes (millions of events and thousands of attributes are now of common usage). SLT is applied successfully in various fields: finance, bioinformatics, marketing, with an obvious and strong intersection with what is called the Data Mining process.

This special issue addresses both theoretical and applied aspects of SLT and its interfaces with statistics, multidimensional classical methods and specific applied fields. Its main purpose—“hopefully”—is to show how many bridges already exist between classical statistics and SLT, a powerful source to cross-fertilization for future research. The issue features nine papers briefly described below.

Those papers were first presented in a November 2002 conference, ‘Statistical Learning, Theory and Applications’, held at Amphithéâtre Grégoire at Conservatoire National des Arts et Métiers (CNAM) Paris, where more than 250 attendees came. The conference was co-sponsored by CNAM, Société Française de Statistique (SFdS), and KXEN Inc.

In *Supervised classification and tunnel vision*, David Hand analyses the dilemma that users of complex SLT methods for supervised classification meet: maintain interpretability while achieving an excellent performance in terms of misclassification rate. He provides an illustration for a new method to solve this dilemma.

The use of Kernel based geometry has now become a major tool for SLT: in *A Tutorial on v-Support Vector Machiness*, Pai-Hsuen Chen, Chih-Jen Lin and Bernard Schölkopf describe the main ideas of statistical learning theory, support vector machines and kernel feature spaces with a particular emphasis for the so-called v-SVM.

Very large scale learning problems remain mainly unsolved: for instance learning how to recognize objects in arbitrary scenes using TV broadcasts as a data source. Current learning algorithms do not scale well enough to allow a timely treatment of such massive data. In *Learning VERY large data sets*, Leon Bottou and Yann Le Cun show that simple on-line

algorithms converge to a ‘true’ solution of the learning problem as fast (asymptotically) as any other algorithm.

Wavelets theory can be approached, partly because of its regularization based methodologies, through the perspective of SLT: in *Wavelet Kernels in RKHS*, A. Rakotomamonjy, X. Mary and S. Canu show a method for building a RKHS spanned by wavelets, and give a few toy-data examples for illustration.

Building models where attributes are functions has been solved in classical multivariate statistics. In *statistical modelling of functional data*, P. Besse, H. Cardot, R. Faivre, M. Goulard give an example of robustification, through smoothing and regularization, with a real case analysis on teledetection.

In *Support vector learning for frontier estimation*, G. Bouchard and S. Girard discuss a methodology to estimate, through a Parzen–Rosenblatt kernel based estimator, the edge or frontier of a set S , given a random set of points drawn from the interior. The regularization factor comes from the Kernel bandwidth, and by analogies with support vector machine theory, a subset of relevant points that define the set S frontier are obtained. A simulated example illustrates the method.

In *A statistical approach for separability of classes*, D. Zighed, S. Lallich and F. Muhlenbach discuss a geometrical graph based methodology—cut edge weight non-parametrical test—to test separability and build class separation in a set of points in R^p . Their method is tested on 13 benchmarks from UCI Repository.

Multiclass discrimination has always been a hot topic, and SVM approach to it (M-SVM) keeps its research potential intact; In *Bound on the risk for M-SVM*, Y. Gueurneur, A. Elisseeff and D. Zelus set up a theoretical framework to compare which of the two main methodologies—methodologies, that is to say a set of dichotomies or a straightforward solution to the m -class problem—is best for generalization.

Kriging is a perfect example (with splines and wavelets) of superb theories that today bridge strongly with SLT on one side, Statistics on the other side. In *Intrinsic kriging and prior information*, E. Vazquez and E. Walter describe such bridges, through RKHS theory—in kriging, the kernel is seen as a covariance—that also apply to RBF, splines, ridge regression and SVM. They also show how intrinsic kriging, a semi-parametric expression of kriging, can be used to introduce effective a priori information, with a application for flow analysis in water pipes.

GILBERT SAPORTA

Chair of Applied Scientist CNAM, Paris, France

MICHEL BERA

Co-Founder and Chief Scientist of KXEN INC