European Network on New Sensing Technologies for Air Pollution Control and Environmental Sustainability - *EuNetAir*

COST Action TD1105

4th International Workshop *EuNetAir* on

Innovations and Challenges for Air Quality Control Sensors

FFG - Austrian Research Promotion Agency - Austrian COST Association

Vienna, Austria, 25 - 26 February 2016

COMPLEXITY OF SOFTCOMPUTING MODELS FOR BIG DATA PROCESSING

Věra Kůrková
MC Member
vera@cs.cas.cz

Roman Neruda
MC Substitute
roman@cs.cas.cz

Institute of Computer Science, Czech Academy of Sciences, Czech Republic

COST is supported by the EU Framework Programme

ESF provides the COST Office through a European Commission contract

Sensor Data

Sensor data sets are growing rapidly in part because they are increasingly gathered by cheap and numerous sensors

large scale, high-dimensionality, incompleteness, and noisy character
expensive labeling

Management is not only about storage and access to data, but analytics plays an important role in making sense of data and exploiting their value

Big data

Volume – large quantity of generated and stored date
Velocity – high speed of increase of volume of data
Veracity – inconsistence, incompleteness, noisy

Learning from data

Labeled data can be used as **training sets** for supervised learning of various soft-computing models such as **neural, radial or kernel networks** with a **capability of generalization**

Generalization = capability of a computational model to satisfactorily process data that were not used for training

Labeling of sensor data is expensive, smaller labeled sets can be used for training of suitable networks

Properly trained networks can predict outputs for unlabeled or new data, estimate missing data

Machine Learning

Learning from data

empirical data

$\{(u_i, v_i) \mid i = 1, \ldots, m\}$

training set = sample of input/output pairs

conceptual data

some global property of input/output function

(smoothness, localization)

Learning is an Optimization Task

optimal parameters of computational models are searched by learning algorithms minimizing error functionals

$$z = \{(u_i, v_i) \,|\, i = 1, \ldots, m\} \subseteq \mathbb{R}^d \times \mathbb{R}$$

sample of data (input/output pairs)

Empirical error functional

$$\mathcal{E}_z(f) = \frac{1}{m} \sum_{i=1}^{m} (f(u_i) - v_i)^2$$

Learning with Generalization

Tikhonov regularization

$$\mathcal{E}_z + \gamma\Psi$$

Ψ functional modeling conceptual data by penalizing some undesired property of input/output functions, e.g., high-frequency oscillations $\Psi(f) = \int \frac{\tilde{f}(s)^2}{\tilde{k}(s)}\, ds \qquad \lim_{\|s\|\to\infty} \frac{1}{\tilde{k}(s)} = \infty$

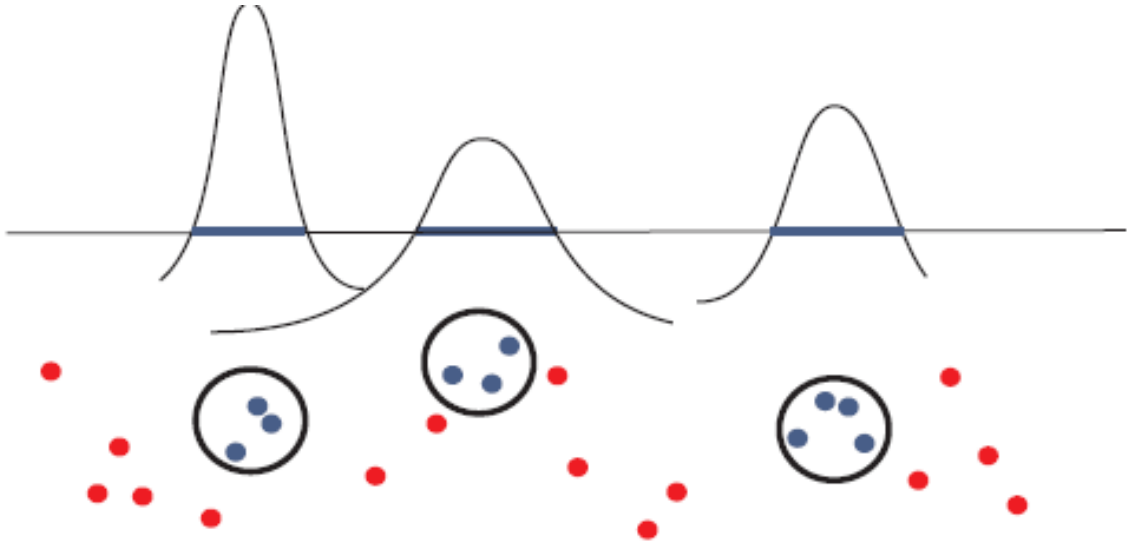
γ trade-off between empirical and conceptual data

Kernel Networks

Networks with computational units defined by positive definite kernel functions (paradigmatic example is Gaussian)

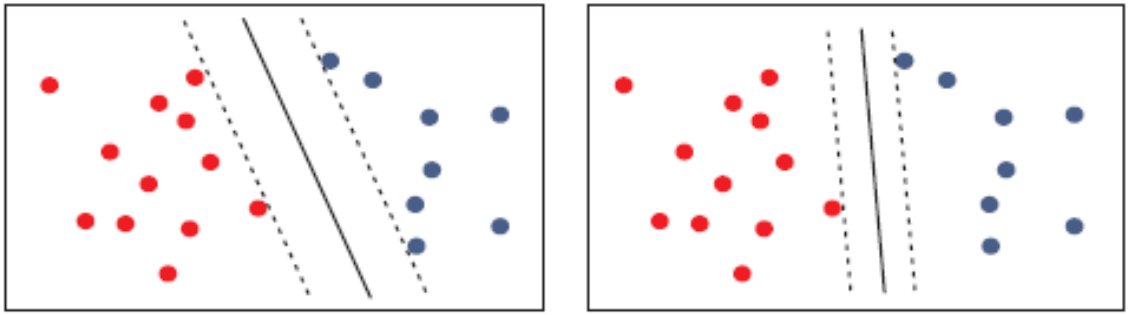
provide: - good stabilizers (defined by norms induced by kernels)

- embedings of data into higher dimensional spaces where more sets of data are linearly separable (support vector machine algorithm)



Generalization by Kernel Networks

Algorithms for kernel networks achive good generalization by maximization of margins between different classes of data



Challenge - Big Data

Many algorithms whose objective is good generalization

do not scale well !

They achieve good generalization on the account of network complexity

Curse of dimensionality = numbers of network units grow exponentially with the input dimension

or

numbers of units are equal to sizes of training data

Multiobjective Optimization

Two objectives: sparsity and generalization

We proposed a **new measure of sparsity prior** of networks in terms of stabilizers defined as

variational norms induced by types of computational units

Regularization with stabilizers in the form of variational norms **reduces model complexity** and **improves generalization** by reducing oscillations (high variations of solutions)

easily implementable - **variational norms can be decreased by output-weight decay regularization techniques**

Concrete learning algorithms

we proposed hybrid evolutionary learning algorithms, implemented, and tested on

data on air chemical compounds from DeVito et al. (tens of thousands of measurements of gas multi-sensor devices recording several air pollutants collocated with conventional air pollution monitoring stations that provide labels for the data)

because of heterogenous character of data we used

composite kernels (in the form od sums and tensor products operating on subsets of input variables)

good mathematical properties suitable for regularization

CONCLUSIONS

We investigated possibilities of supervised learning from large sensor data sets

We derived theoretical results on sparsity and generalization of computational models and proposed composite kernels computational units suitable for heterogenous sensor data

We designed hybrid learning algorithms for composite kernel networks and tested them on sensor data from colleagues from COST action EuNetAir