Semi-Supervised Learning of Semantic Spatial Concepts for a Mobile Robot

Nicolò Cesa-Bianchi[1] and Francesco Orabona[2]

[1]Università degli Studi di Milano

[2]Toyota Technological Institute, Chicago

Summary

1. Project description
2. Online multikernel learning
3. Selective sampling

Summary

1. Project description
2. Online multikernel learning
3. Selective sampling

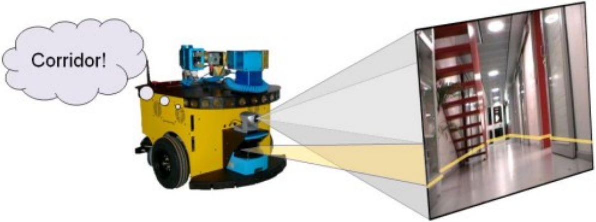
Project details

- Participants:
  Nicolò Cesa-Bianchi (Milano) and Barbara Caputo (Martigny)
- Duration: 18 months
- Starting date: January 1, 2010
- Budget: 55,100 Euros
- Kick-off meeting: February 7-8, 2010 in Martigny

Note

Due to hiring problems, only the first 12 months of the project have been carried out so far

The place recognition problem in robot navigation



- The ability of building robust semantic space representations of environments is crucial for the development of truly autonomous robots
- Online learning is important in dynamic environments
- Semisupervised/active learning provides a realistic interaction protocol

Project's main goals

- ☺ Online algorithms for sparse multiview/multikernel learning
- ☺ Advancing state-of-the-art in selective sampling (online active learning)
- ☺ Selective sampling algorithms based on non-Euclidean norms
- Integration and refinement of approaches
- Testing on robotic platform

Publications

N. Cesa-Bianchi, C. Gentile, and F. Orabona
Robust bounds for classification via selective sampling
ICML 2009

L. Jie, F. Orabona, M. Fornoni, B. Caputo, and N. Cesa-Bianchi
OM-2: An online multi-class multi-kernel learning algorithm
4th IEEE Online Learning for Computer Vision Workshop, 2010

F. Orabona and N. Cesa-Bianchi
Better algorithms for selective sampling
ICML 2011

Summary

1. Project description
2. Online multikernel learning
3. Selective sampling

Online multikernel learning

Online integration of data coming from robot sensors

Kernel-based multiview learning

1. Fix set $\mathcal{H}_1, \ldots, \mathcal{H}_N$ of RKHS
2. Simultaneous online learning of N models $f_1 \in \mathcal{H}_1, \ldots, f_N \in \mathcal{H}_N$
3. Online prediction using combination of models

$$f(x) = \sum_{i=1}^{N} \alpha_i f_i(x) \qquad \alpha_i \in \mathbb{R}$$

Baselines:

- Single best trained model in $\{f_1, \ldots, f_N\}$
- Flat margin average $\quad \dfrac{1}{N} \sum_{i=1}^{N} f_i(x)$

The mother of all online algorithms

Online Mirror Descent

Parameters: Regularization function R and learning rate $\eta > 0$

Initialize: $\boxed{g = \mathbf{0}}$ `// primal parameter`

For each data item in the stream

1. Predict label using $f = \nabla R^*(g)$ `// mirror step`
2. Observe true label and suffer loss $\ell_t(f)$
3. Update $\boxed{g \leftarrow g - \eta \nabla \ell_t(f)}$ `// gradient step`

Remarks

- Regularizer R is any strongly convex function (e.g., squared norm)
- Performance bounds good when data match choice of regularizer (e.g., sparsity)
- f, g can belong to any pair of dual linear spaces, such as matrices (then R is typically a squared matrix norm)

Online multikernel learning (cont.)

Multikernel group-norm Perceptron

OMD with squared $(2, p)$ matrix group norm as regularizer

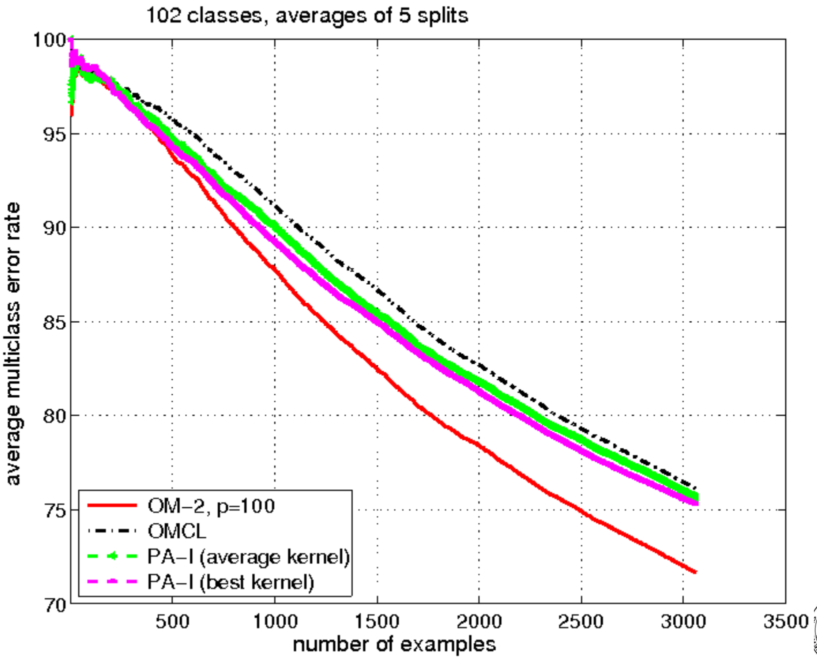
$$R\Big(\big[f_1, \dots, f_N\big]\Big) = \Big\|\big(f_1, \dots, f_N\big)\Big\|_{2,p}^2 = \Big|\big(\|f_1\|_2, \dots, \|f_N\|_2\big)\Big|_p^2$$

- For $p = 2$, algorithm does flat margin average, $\alpha_i = \frac{1}{N}$
- For $p > 2$, performance improves when $\Big(\|f_1^*\|_2, \dots, \|f_N^*\|_2\Big)$ is a sparse vector, where $f_1^*, \dots, f_N^*$ are the "best" models in each RKHS

Datasets

Caltech-101

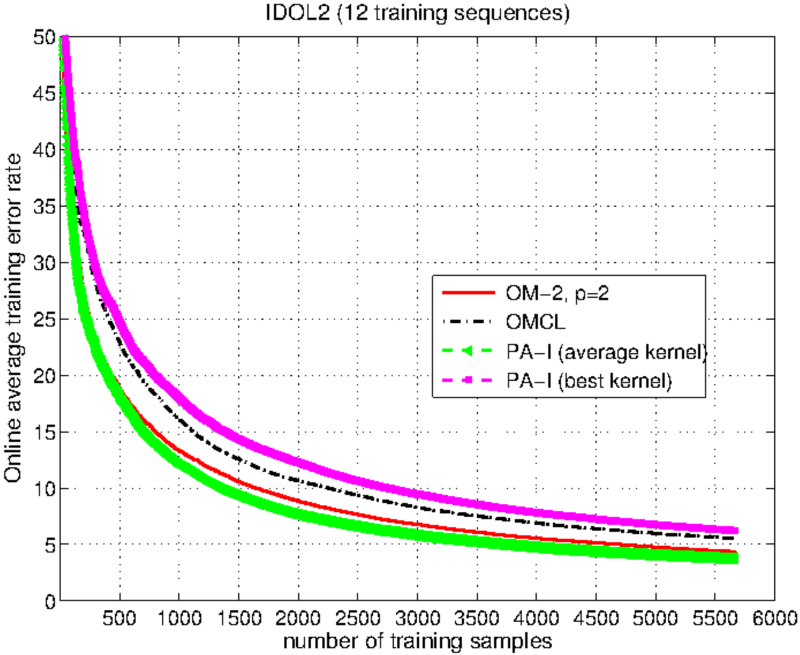
- Standard benchmark dataset for object categorization
- 102 classes, 3K images
- 48 kernels


102 classes, averages of 5 splits
average multiclass error rate
number of examples
100
95
90
85
80
75
70
500
1000
1500
2000
2500
3000
3500
OM−2, p=100
OMCL
PA−I (average kernel)
PA−I (best kernel)

Datasets

Subset of IDOL2

- 12 image sequences acquired using a perspective camera mounted on a mobile robot platform
- Sequences captured in an indoor laboratory environment consisting of five different rooms under various weather and illumination conditions
- 5 classes (rooms), 6K images, 4 views (3 visual, 1 laser scan)


IDOL2 (12 training sequences)
Online average training error rate
number of training samples
0
5
10
15
20
25
30
35
40
45
50
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
OM−2, p=2
OMCL
PA−I (average kernel)
PA−I (best kernel)

Summary

Making each label more worth

- Family $\mathcal{F}$ of real functions $f : \mathbb{R}^d \to \mathbb{R}$
- Associated binary classifiers $\textsc{sgn}(f) : \mathbb{R}^d \to \{-1, +1\}$

Standard supervised learning (full sampling)

- In the training phase, query human expert t consecutive times
- Then choose $\widehat{f}_t \in \mathcal{F}$ based on sample

$$(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_t, Y_t) \in \mathbb{R}^d \times \{-1, +1\}$$

Question

Can we exploit labels better by subsampling adaptively the label process?

Data model

Ingredients

- RKHS $\mathcal{H}$
- Stochastic label process $\mathbb{E}\big[Y_t \mid \boldsymbol{x}_t\big] = f^*(\boldsymbol{x}_t)$ for some $f^* \in \mathcal{H}$
- Deterministic data process $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots$ s.t. $\max_t |f^*(\boldsymbol{x}_t)| \leqslant 1$
- If $\mathcal{H}$ is universal, then it can approximate any label process

Regret

$$R_T = \frac{1}{T}\sum_{t=1}^{T} \mathbb{P}\big(Y_t\, \widehat{f}_t(\boldsymbol{x}_t) \leqslant 0\big) - \frac{1}{T}\sum_{t=1}^{T} \mathbb{P}\big(Y_t\, f^*(\boldsymbol{x}_t) \leqslant 0\big)$$

The BBQ algorithm

RLS estimate (primal form for simplicity)

$$\widehat{f}_t(\boldsymbol{x}_t) = \boldsymbol{x}_t^\top \big(\underbrace{I + S\,S^\top + \boldsymbol{x}_t \boldsymbol{x}_t^\top}_{A} \big)^{-1} S\,\boldsymbol{Y}$$

$S = \big[\boldsymbol{x}_{t_1}, \ldots, \boldsymbol{x}_{t_N}\big]$ matrix of queried points

$\boldsymbol{Y} = \big(Y_{t_1}, \ldots, Y_{t_N}\big)$ vector of queried labels

BBQ query rule: If $\boldsymbol{x}_t^\top A^{-1} \boldsymbol{x}_t > t^{-\frac{1}{\kappa}}$ then query $\boldsymbol{x}_t$ $(\kappa \geqslant 1)$

Intuition: $\boldsymbol{x}_t^\top A^{-1} \boldsymbol{x}_t$

1. is large when $\boldsymbol{x}_t$ is not correlated with any principal component of the past queried data
2. is an upper bound on both bias and variance of $\widehat{f}_t(\boldsymbol{x}_t)$ w.r.t. the label process

Rates for BBQ

Theorem

$$R_T \leqslant \min_{0<\varepsilon<1} \left[\varepsilon\, T_\varepsilon + \frac{1+\|f^*\|^2}{T\varepsilon\kappa} \ln(T) \ln|A_{T+1}| + \frac{1}{T\varepsilon^{2\kappa}} \left(\lceil \kappa \rceil ! + \|f^*\|^{2\kappa} \right) \right]$$

$$N_T \leqslant \frac{\ln|A_{T+1}|}{T^{1-\frac{1}{\kappa}}}$$

T_ε *is fraction of points* x_t *such that* $\left|f^*(x_t)\right| \leqslant \varepsilon$

Case $d < \infty$

Under standard Tsybakov noise condition, regret-per-query is

$$R_N \leqslant \left(\frac{d(\ln N)^2}{N} \right)^{\frac{1+\alpha}{2}}$$

This rate is optimal, but BBQ needs to know α

Other RLS-based selective samplers

DGS-mod query rule [Dekel, Gentile, Sridharan, 2010]

If $\quad \widehat{f}_t(\boldsymbol{x}_t)^2 < (\kappa \ln t)\big(\boldsymbol{x}_t^\top A^{-1}\boldsymbol{x}_t\big)\big[\ldots\big] \quad$ then query $\boldsymbol{x}_t$

- Regret and query bounds incomparable to those of BBQ
- Under Tsybakov condition, regret-per-query rate matches that of A^2 without need of knowing α

SS query rule: If $\quad \widehat{f}_t(\boldsymbol{x}_t)^2 \leqslant \dfrac{\kappa \ln t}{N_t} \quad$ then query $\boldsymbol{x}_t$

SOLE query rule: Query x_t with probability $\quad \dfrac{1}{1+\kappa\big|\widehat{f}_t(\boldsymbol{x}_t)\big|}$

Active multiview online learning

- RLS-based algorithms work only with Euclidean regularization
- Need OMD-based semisupervised algorithms with:
  1. Simultaneous bounds on regret and query rate
  2. Regret still controlled by fit of regularizer on data

A hybrid algorithm

DGS-hybrid

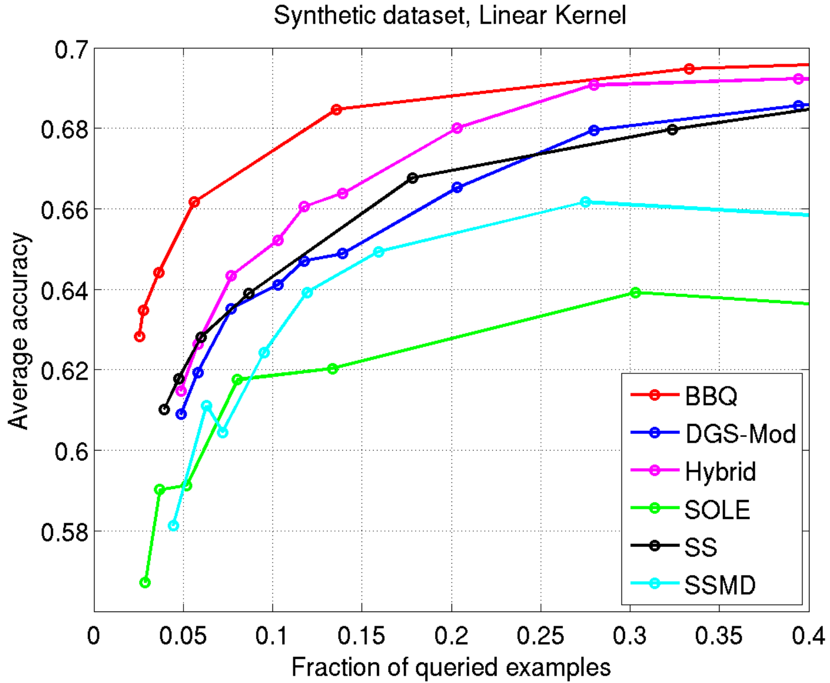
For each data item in the stream

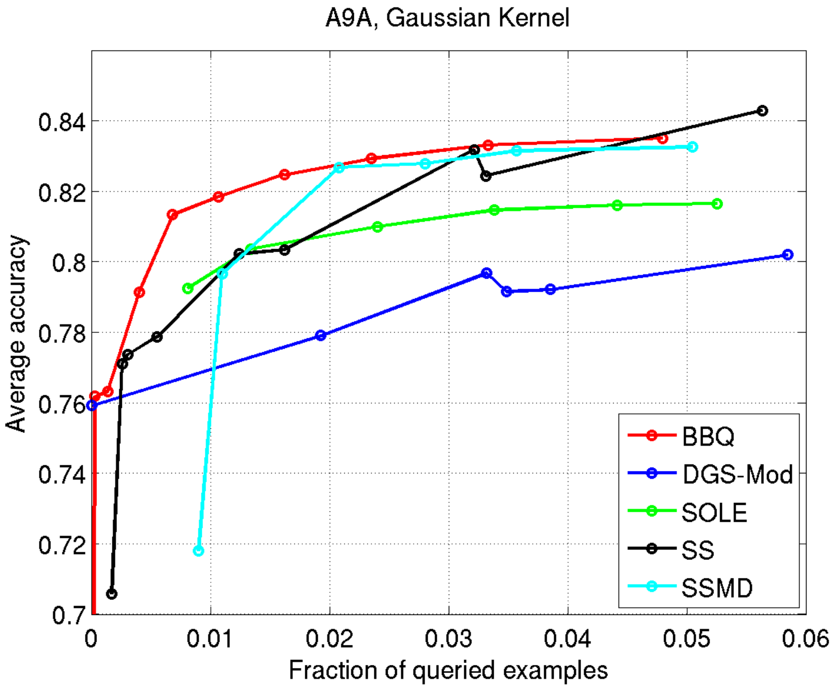
1. Compute current models for DGS-mod and OMD
2. If DGS-mod wants to query, then predict with OMD and make the query, use label to update both models
3. Otherwise, predict with DGS-mod

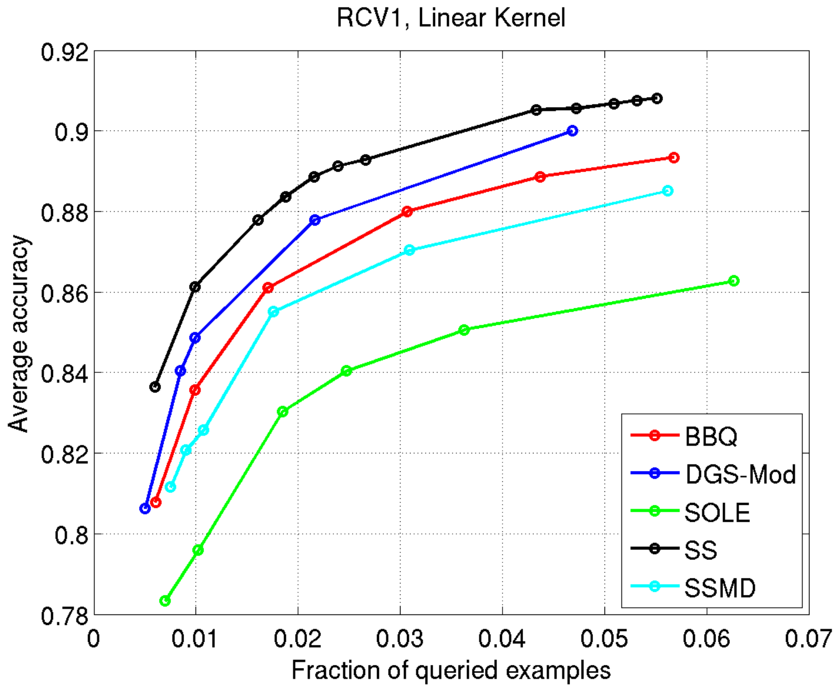
Results

$$\mathbb{E}\left[\text{Mistakes}\right] \leqslant \inf_{f}\left(\text{HingeLoss}(f) + c\sqrt{R(f)\,N_T}\right) + \mathcal{O}(1)$$

$$\mathbb{E}\left[\text{Queries}\right] = N_T \qquad \text{(DGS-mod query rate)}$$


Synthetic dataset, Linear Kernel
Average accuracy
Fraction of queried examples
0.7
0.68
0.66
0.64
0.62
0.6
0.58
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
BBQ
DGS-Mod
Hybrid
SOLE
SS
SSMD


A9A, Gaussian Kernel
Average accuracy
Fraction of queried examples
0.7
0.72
0.74
0.76
0.78
0.8
0.82
0.84
0
0.01
0.02
0.03
0.04
0.05
0.06
BBQ
DGS-Mod
SOLE
SS
SSMD


RCV1, Linear Kernel
Average accuracy
Fraction of queried examples
0.92
0.9
0.88
0.86
0.84
0.82
0.8
0.78
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
BBQ
DGS-Mod
SOLE
SS
SSMD

Conclusions

- Online multiview learning based on matrix group norms
- Better understanding of RLS-based selective sampling algorithms
- First attempt to design non-Euclidean active algorithms with simultaneous regret/query performance guarantees