Robust Bounds for Classification via Selective Sampling

Nicolò Cesa-Bianchi Claudio Gentile
Francesco Orabona

DSI, Università degli Studi di Milano, Milano, Italy

DICOM, Università dell'Insubria, Varese, Italy

Idiap Research Institute, Martigny, Switzerland

International Conference on Machine Learning 2009

Active Learning

- What? Active learning algorithms, *selecting* the data to be labeled, can achieve a significant boost over batch learning.
- Why? Unlabeled data is cheap, labeling is expensive.
- But! Most previous studies consider the case when instances are drawn i.i.d. from a fixed distribution.

ICML09 Tutorial on Active Learning (S. Dasgupta and J. Langford)

Future work for all of us

1. Foundations Is active learning possible in a fully adversarial setting?
2. Application Is an active learning reduction to supervised possible without constraints?
3. Extension What about other settings for interactive learning? (structured? partial label? Differing oracles with differing expertise?)
4. Empirical Can we achieve good active learning performance with a consistent algorithm on a state-of-the-art problem?

Further discussion at http://hunch.net

ICML09 Tutorial on Active Learning (S. Dasgupta and J. Langford)

Future work for all of us

1. Foundations Is active learning possible in a fully adversarial setting?
2. Application Is an active learning reduction to supervised possible without constraints?
3. Extension What about other settings for interactive learning? (structured? partial label? Differing oracles with differing expertise?)
4. Empirical Can we achieve good active learning performance with a consistent algorithm on a state-of-the-art problem?

Further discussion at http://hunch.net

Outline

Outline

1. Problem definition
   - Previous work
   - Hypothesis
2. Bound on Bias Query
   - Regularized Least Square
   - BBQ
   - Parametric BBQ
3. Experimental Results

Selective Sampling

- Selective sampling is a well-known semi-supervised online learning setting [CAL90].
- At each step $t = 1, 2, \ldots$ the learner receives an instance $\boldsymbol{x}_t \in \mathbb{R}^d$ and outputs a binary prediction for the associated unknown label $y_t \in \{-1, +1\}$.
- After each prediction the learner may observe the label y_t only by issuing a *query*. If no query is issued at time t, then y_t remains unknown.
- Since one expects the learner's performance to improve if more labels are observed, our goal is to trade off predictive accuracy against number of queries.
- No i.i.d. hypothesis!

Selective Sampling

- Selective sampling is a well-known semi-supervised online learning setting [CAL90].
- At each step $t = 1, 2, \ldots$ the learner receives an instance $\boldsymbol{x}_t \in \mathbb{R}^d$ and outputs a binary prediction for the associated unknown label $y_t \in \{-1, +1\}$.
- After each prediction the learner may observe the label y_t only by issuing a *query*. If no query is issued at time t, then y_t remains unknown.
- Since one expects the learner's performance to improve if more labels are observed, our goal is to trade off predictive accuracy against number of queries.
- No i.i.d. hypothesis!

Previous Work

- Most previous studies consider the case when instances are drawn i.i.d. from a fixed distribution.
- Some exception:
  - The work [CGZ06] is completely worst case, however, they are unable prove bounds on the label query rate.
  - In the KWIK model of [SL08,LLW08] the goal is to approximate the Bayes margin to within a given accuracy ε. It assumes arbitrary sequences of instances and a linear stochastic model for labels. Their algorithm competes against an adaptive adversarial strategy for generating instances, by asking $\widetilde{\mathcal{O}}(d^3/\varepsilon^4)$ queries.
- We consider a setting similar to the KWIK one.

Hypothesis: Label Noise Model

- All results proven hold for *any fixed individual sequence* $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ of instances, $\boldsymbol{x}_t \in \mathbb{R}^d$, under the sole assumption that $\|\boldsymbol{x}_t\| = 1$ for all $t \geq 1$.
- We assume the corresponding labels $y_t \in \{-1, +1\}$ are realizations of random variables Y_t such that $\mathbb{E}\, Y_t = \boldsymbol{u}^\top \boldsymbol{x}_t$ for all $t \geq 1$, where $\boldsymbol{u} \in \mathbb{R}^d$ is a fixed and unknown vector such that $\|\boldsymbol{u}\| = 1$.
  - Note that $\text{SGN}(\Delta_t)$, for $\Delta_t = \boldsymbol{u}^\top \boldsymbol{x}$, is the Bayes optimal classifier for this noise model.
  - This noise model can be made highly nonlinear via kernel functions.

Outline

1. Problem definition
   - Previous work
   - Hypothesis
2. Bound on Bias Query
   - Regularized Least Square
   - BBQ
   - Parametric BBQ
3. Experimental Results

Our Main Tool: Regularized Least Square

- Our algorithms are based on RLS.
- We use a well-known variant RLS estimate, that can be efficiently run in any RKHS,

$$\boldsymbol{w}_t = \left(I + S_{t-1}\, S_{t-1}^\top + \boldsymbol{x}_t \boldsymbol{x}_t^\top\right)^{-1} S_{t-1}\, \boldsymbol{Y}_{t-1}$$

defined over the matrix $S_{t-1} = \left[\boldsymbol{x}'_1, \dots, \boldsymbol{x}'_{N_{t-1}}\right]$ of the N_{t-1} queried instances up to time $t-1$. The random vector $\boldsymbol{Y}_{t-1} = \left(Y'_1, \dots, Y'_{N_{t-1}}\right)$ contains the observed labels.
- Note that the current sample $\boldsymbol{x}_t$ is included in the formula.

Bound on Bias Query: the BBQ Algorithm

Parameters: $0 \leq \kappa \leq 1$
for each time step $t = 1, 2, \dots$ **do**
  Observe instance $\boldsymbol{x}_t \in \mathbb{R}^d$
  $\widehat{\Delta}_t = \boldsymbol{w}_t^\top \boldsymbol{x}_t$ (RLS)
  predict label $y_t \in \{-1, +1\}$ with $\text{SGN}(\widehat{\Delta}_t)$
  $r_t = \boldsymbol{x}_t^\top \left(I + S_{t-1} S_{t-1}^\top + \boldsymbol{x}_t \boldsymbol{x}_t^\top\right)^{-1} \boldsymbol{x}_t$
  if $r_t > t^{-\kappa}$ **then**
    query label y_t
  end if
end for

- Kernels can be used, formulating the algorithm in dual variables.
- The space and time complexity to predict and update is $\mathcal{O}\left(d^2\right)$ for the primal version and $\mathcal{O}\left(N_{t-1}^2\right)$ for the dual version.

Regret Bound for BBQ

Theorem

If BBQ is run with input $\kappa \in [0,1]$ *then its cumulative regret* $R_T = \sum_{t=1}^{T} \left(\mathbb{P}(Y_t \widehat{\Delta}_t < 0) - \mathbb{P}(Y_t \Delta_t < 0) \right)$ *after any number T of steps satisfies*

$$R_T \leq \min_{0<\varepsilon<1} \left(\varepsilon\, T_\varepsilon + \mathcal{O}\left(\frac{1}{\varepsilon^{2/\kappa}} + \frac{d}{\varepsilon^2} \ln T \right) \right) ,$$

where $T_\varepsilon = \left| \{ 1 \leq t \leq T \, : \, |\Delta_t| < \varepsilon \} \right|$.
The number of queried labels is $N_T = \mathcal{O}\left(d\, T^{\kappa} \ln T \right)$.

The BBQ Algorithm

Parameters: $0 \le \kappa \le 1$
for each time step $t = 1, 2, \dots$ **do**
  Observe instance $\boldsymbol{x}_t \in \mathbb{R}^d$
  $\widehat{\Delta}_t = \boldsymbol{w}_t^\top \boldsymbol{x}_t$ (RLS)
  predict label $y_t \in \{-1, +1\}$ with $\text{SGN}(\widehat{\Delta}_t)$
  $r_t = \boldsymbol{x}_t^\top \left(I + S_{t-1} S_{t-1}^\top + \boldsymbol{x}_t \boldsymbol{x}_t^\top\right)^{-1} \boldsymbol{x}_t$
  if $r_t > t^{-\kappa}$ **then**
    query label y_t
  end if
end for

r_t is related the "distance of the current sample from the queried samples".

How Does It Work?

- BBQ issues a query when a common upper bound on bias and variance of the current RLS estimate is larger than a given threshold.
- The bound depends on r_t.
- When this upper bound gets small, we infer via a large deviation argument that the margin of the RLS estimate on the current instance is close enough to the margin of the Bayes optimal classifier.
- Hence the learner can safely avoid issuing a query on that step.
- r_t does not depend on the labels, similarly to [SL08].

How Does It Work?

- BBQ issues a query when a common upper bound on bias and variance of the current RLS estimate is larger than a given threshold.
- The bound depends on r_t.
- When this upper bound gets small, we infer via a large deviation argument that the margin of the RLS estimate on the current instance is close enough to the margin of the Bayes optimal classifier.
- Hence the learner can safely avoid issuing a query on that step.
- r_t does not depend on the labels, similarly to [SL08].

What Happens If We Know ε?

- Most technicalities in the proof are due to the fact that the final bound depends on the optimal choice of this ε, which the algorithm need not know.
- Suppose we want to approximate the Bayes margin to a given precision ε.
- We can design an algorithm that
  - when it does not query the label, it gives a prediction that is within an error of ε to the Bayes margin with high probability;
  - number of queries *logarithmic* in time.

The Parametric BBQ Algorithm

Parameters: $0 < \varepsilon, \delta < 1$
for each time step $t = 1, 2, \ldots$ **do**
  observe instance $\boldsymbol{x}_t \in \mathbb{R}^d$
  $\widehat{\Delta}_t = \boldsymbol{w}_t^\top \boldsymbol{x}_t$ (RLS)
  predict label $y_t \in \{-1, +1\}$ with $\mathrm{SGN}(\widehat{\Delta}_t)$
  $r_t = \boldsymbol{x}_t^\top \left(I + S_{t-1} S_{t-1}^\top + \boldsymbol{x}_t \boldsymbol{x}_t^\top\right)^{-1} \boldsymbol{x}_t$
  $q_t = S_{t-1}^\top \left(I + S_{t-1} S_{t-1}^\top + \boldsymbol{x}_t \boldsymbol{x}_t^\top\right)^{-1} \boldsymbol{x}_t$
  $s_t = \left\| \left(I + S_{t-1} S_{t-1}^\top + \boldsymbol{x}_t \boldsymbol{x}_t^\top\right)^{-1} \boldsymbol{x}_t \right\|$
  if $\left[\varepsilon - r_t - s_t\right]_+ < \|\boldsymbol{q}_t\| \sqrt{2 \ln \frac{t(t+1)}{2\delta}}$ **then**
    query label y_t
  end if
end for

Regret Bound of Parametric BBQ

Theorem

If Parametric BBQ is run with input $\varepsilon, \delta \in (0, 1)$ *then:*

- *with probability at least* $1 - \delta$, $\left|\widehat{\Delta}_t - \Delta_t\right| \leq \varepsilon$ *holds on all time steps* t *when no query is issued;*
- *the number* N_T *of queries issued after any number* T *of steps is bounded as*

$$N_T = \mathcal{O}\left(\frac{d}{\varepsilon^2}\left(\ln\frac{T}{\delta}\right)\ln\frac{\ln(T/\delta)}{\varepsilon}\right) .$$

Is It Possible To Obtain a Better Bound?

Is It Possible To Obtain a Better Bound?

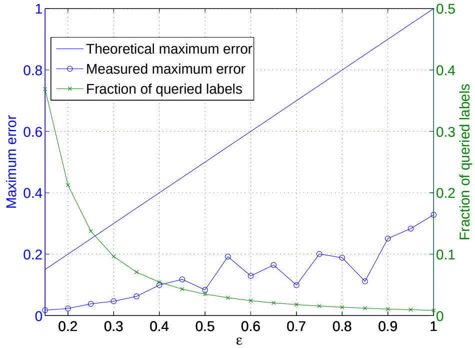
No!

- The bound on the number of queried labels is optimal up to logarithmic factors!
- At least $\Omega(d/\varepsilon^2)$ queries are needed to learn any target hyperplane with arbitrarily small accuracy and arbitrarily high confidence.

Outline

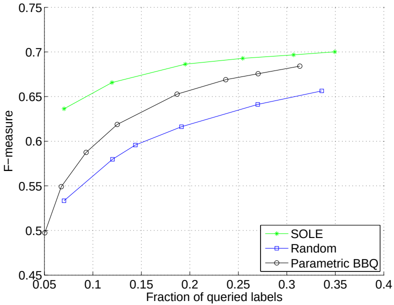
1. Problem definition
   - Previous work
   - Hypothesis
2. Bound on Bias Query
   - Regularized Least Square
   - BBQ
   - Parametric BBQ
3. Experimental Results

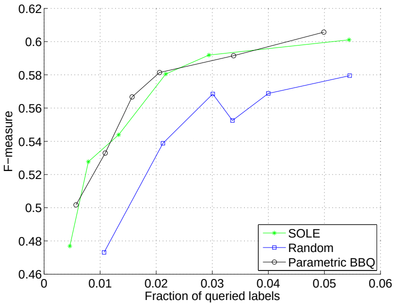
Synthetic Experiment



- We tested Parametric BBQ.
- 10,000 random examples on the unit circle in $\mathbb{R}^2$.
- The labels were generated according to our noise model using a randomly selected hyperplane $\boldsymbol{u}$ with unit norm.

Real World Experiments





F-measure and fraction of queried labels for different algorithms on Adult9 dataset (left)(Gaussian Kernel) and RCV1 (right)(linear kernel).

Summary

- We have introduced a new family of online algorithms, the BBQ family, for selective sampling under (oblivious) adversarial environments.
- These algorithms naturally interpolate between fully supervised and fully unsupervised learning. Parametric BBQ is designed to work in a weakened KWIK framework with improved bounds on the number of queried labels.

Work in Progress

- Extending the algorithms to work with adaptive adversary.
- Improving the bound on the number of queried labels, removing the logarithmic dependency on the time.

Thanks for your attention