Combinatorial Approach to Generalization Bounds Tightening

Konstantin Vorontsov
(voron@ccas.ru, www.ccas.ru/voron)

Computing Centre of Russian Academy of Sciences,
Vavilova 40, 119991, Moscow, Russian Federation

7th Open German/Russian Workshop (OGRW-7)
on Pattern Recognition and Image Understanding
August 20–23, 2007
Ettlingen, Germany

Outline

Dear Colleagues, I shall speak about one of the most challenging problem in computational learning theory — the problem of generalization ability of learning algorithms. I shall start from a general theoretical framework, then consider a fundamental problem of learning theory — the looseness of generalization bounds. I shall finish by some empirical results.

Weak Probability Axioms

1. $X^L = \{x_i\}_{i=1}^L$ — a given finite subset from a set of objects $\mathbb{X}$.
2. All partitions $X^L = X_n^\ell \cup X_n^k, \ \ n = 1, \ldots, N$, where $N = C_L^k, \ \ L = \ell + k$, are equally probable.

Then...

Let us start from two very simple axioms. First, [see...] we assume that in data analysis one can observe only a finite set of objects X^L. The set can be unknown, but it never can be infinite.

Second, [see...] we assume that objects appear at random, so that all partitions of our set into two subsets have equal chances to realize. The order of objects is the only source of randomness in our framework.

Weak Probability Axioms

1. $X^L = \{x_i\}_{i=1}^L$ — a given finite subset from a set of objects $\mathbb{X}$.
2. All partitions $X^L = X_n^\ell \cup X_n^k$, $n = 1, \ldots, N$, where $N = C_L^k$, $L = \ell + k$, are equally probable.

Then

- Consider an event A as a function $A\colon \{1, \ldots, N\} \to \{0, 1\}$
- The fraction of partitions n: $A(n) = 1$ can be interpreted as probability or expectation:

$$\mathsf{P}_n A(n) \equiv \mathsf{E}_n A(n) = \frac{1}{N} \sum_{i=1}^{N} A(n).$$

Here "probability" P_n is simply averaging operator $\frac{1}{N}\sum_{i=1}^{N}$.

Weak Probability Axioms

1. $X^L = \{x_i\}_{i=1}^L$ — a given finite subset from a set of objects $\mathbb{X}$.
2. All partitions $X^L = X_n^\ell \cup X_n^k$, $n = 1, \ldots, N$, where $N = C_L^k$, $L = \ell + k$, are equally probable.

Then

- Consider an event A as a function $A\colon \{1, \ldots, N\} \to \{0, 1\}$
- The fraction of partitions n: $A(n) = 1$ can be interpreted as probability or expectation:

$$\mathsf{P}_n A(n) \equiv \mathsf{E}_n A(n) = \frac{1}{N} \sum_{i=1}^{N} A(n).$$

Here "probability" P_n is simply averaging operator $\frac{1}{N}\sum_{i=1}^{N}$.

Under these assumptions the probability of an event $A(n)$ [see...] is defined as the fraction of partitions n [see...] for which $A(n)$ is true. Note that there is no difference between expectation [see...] and probability [see...] in this framework.

A general problem of Empirical Prediction

- Empirical framework:
  some partition (X_n^ℓ, X_n^k) realizes, $n \in \{1, \ldots, N\}$;
  subsample X_n^ℓ is *observable*,
  subsample X_n^k is *hidden*.
- Given a function $T\colon \mathbb{X}^k \times \mathbb{X}^\ell \to R$
  1. Chose a function $\hat{T}\colon \mathbb{X}^\ell \to R$ such that
  the value $\hat{T}_n = \hat{T}(X_n^\ell)$ predicts the value $T_n = T(X_n^k, X_n^\ell)$.
  2. Estimate prediction accuracy (obtain an upper bound):

$$\mathsf{P}_n\big[d(\hat{T}_n, T_n) > \epsilon\big] \leq \eta(\epsilon),$$

  where $d(\hat{r}, r)$ — discrepancy function, e.g. $d(\hat{r}, r) = |\hat{r} - r|$.

A general problem of Empirical Prediction

- Empirical framework:
  some partition (X_n^ℓ, X_n^k) realizes, $n \in \{1, \ldots, N\}$;
  subsample X_n^ℓ is *observable*,
  subsample X_n^k is *hidden*.
- Given a function $T\colon \mathbb{X}^k \times \mathbb{X}^\ell \to R$
  1. Chose a function $\hat{T}\colon \mathbb{X}^\ell \to R$ such that
  the value $\hat{T}_n = \hat{T}(X_n^\ell)$ predicts the value $T_n = T(X_n^k, X_n^\ell)$.
  2. Estimate prediction accuracy (obtain an upper bound):

$$\mathsf{P}_n\big[d(\hat{T}_n, T_n) > \epsilon\big] \le \eta(\epsilon),$$

  where $d(\hat{r}, r)$ — discrepancy function, e.g. $d(\hat{r}, r) = |\hat{r} - r|$.

The problem of empirical prediction arise when some of the equiprobable partitions [see...] realizes, but one observe only a first subset X_n^ℓ [see...] whereas X_n^k [see...] remains hidden. We want to predict a value of a given function T [see...] that depends on both parts, having an information $\hat{T}$ [see...] computed from the observed part only. Also we want to know in advance how accurate our prediction can be [see...] . Then, the problem is to upper bound the fraction of partitions for which our prediction fails. To formulate exactly what means "fails" we introduce a discrepancy function d [see...] (that can be simply a difference in most cases) and a threshold of exactness ϵ [see...] .

Why we call this framework "Weak Probability Axioms"?

Insight:

Weak PA leads to "simplified Probability Theory"

Indeed...

Insight:
Weak PA leads to "simplified Probability Theory"

Indeed...

Why we call this framework "the weak probability axioms" and even the weak Probability Theory?

Why we call this framework "Weak Probability Axioms"?

Insight:

Weak PA leads to "simplified Probability Theory"

Indeed...

- Weak PA is sufficient to prove fundamental facts:
  - the Law of Large Numbers, with exact convergence rate;
  - Kolmogorov-Smirnov criterion, also exact;
  - many statistical hypothesis tests (order statistics etc.);
  - Vapnik-Chervonenkis generalization bounds (see later);
  - etc.

Insight:
Weak PA leads to "simplified Probability Theory"

Indeed...
- Weak PA is sufficient to prove fundamental facts:
  - the Law of Large Numbers, with exact convergence rate;
  - Kolmogorov-Smirnov criterion, also exact;
  - many statistical hypothesis tests (order statistics etc.);
  - Vapnik-Chervonenkis generalization bounds (see later);
  - etc.

First, because it is very general. Empirical Prediction is one of the central problems in Probability Theory, Statistics and Learning Theory.
In practice predictions are interesting only for finite sets of objects, because nobody can observe an infinite set of objects. Empirical Prediction is transductive by its nature. Many fundamental results in Statistics and Learning Theory can be reformulated in transductive form.

Why we call this framework "Weak Probability Axioms"?

Insight:

Weak PA leads to "simplified Probability Theory"

Indeed...

- Weak PA is sufficient to prove fundamental facts:
  - the Law of Large Numbers, with exact convergence rate;
  - Kolmogorov-Smirnov criterion, also exact;
  - many statistical hypothesis tests (order statistics etc.);
  - Vapnik-Chervonenkis generalization bounds (see later);
  - etc.
- Weak PA is based on less restrictive assumptions:
  - no need of probability measure on $\mathbb{X}$;
  - no need of frequentist definition of probability via $L \to \infty$;

Why we call this framework "Weak Probability Axioms"?

Insight:
Weak PA leads to "simplified Probability Theory"

Indeed...
- Weak PA is sufficient to prove fundamental facts:
  - the Law of Large Numbers, with exact convergence rate;
  - Kolmogorov-Smirnov criterion, also exact;
  - many statistical hypothesis tests (order statistics etc.);
  - Vapnik-Chervonenkis generalization bounds (see later);
  - etc.
- Weak PA is based on less restrictive assumptions:
  - no need of probability measure on $\mathbb{X}$;
  - no need of frequentist definition of probability via $L \to \infty$;

On the other hand, Weak Probability Axioms are less restrictive if compared with classical Kolmogorov's Axioms. Here we don't need of probability measure on object space and we don't define a probability via passage to the limit.

Why we call this framework "Weak Probability Axioms"?

Insight:

Weak PA leads to "simplified Probability Theory"

Indeed...

- Weak PA is sufficient to prove fundamental facts:
  - the Law of Large Numbers, with exact convergence rate;
  - Kolmogorov-Smirnov criterion, also exact;
  - many statistical hypothesis tests (order statistics etc.);
  - Vapnik-Chervonenkis generalization bounds (see later);
  - etc.
- Weak PA is based on less restrictive assumptions:
  - no need of probability measure on $\mathbb{X}$;
  - no need of frequentist definition of probability via $L \to \infty$;
  - no need of "probability" at all!

Why we call this framework "Weak Probability Axioms"?

Insight:
Weak PA leads to "simplified Probability Theory"

Indeed...

- Weak PA is sufficient to prove fundamental facts:
  - the Law of Large Numbers, with exact convergence rate;
  - Kolmogorov-Smirnov criterion, also exact;
  - many statistical hypothesis tests (order statistics etc.);
  - Vapnik-Chervonenkis generalization bounds (see later);
  - etc.
- Weak PA is based on less restrictive assumptions:
  - no need of probability measure on $\mathbb{X}$;
  - no need of frequentist definition of probability via $L \to \infty$;
  - no need of "probability" at all!

Frankly speaking, we don't need of the notion of probability at all. In our framework "probability" is no more than a synonym of "fraction of partitions".

Example: transductive form of the Law of Large Numbers

Def. The *frequency* of $S \subset \mathbb{X}$ in a finite sample $U \subset \mathbb{X}$:

$$\nu_S(U) = \frac{1}{|U|} \sum_{u \in U} [u \in S].$$

Theorem (common knowledge)

The frequency $\nu_S(X_n^\ell)$ predicts the frequency $\nu_S(X_n^k)$.
Prediction accuracy is given by an exact bound

$$\mathsf{P}_n\big[\nu_S(X_n^k) - \nu_S(X_n^\ell) \geq \epsilon\big] = H\binom{\ell}{L}{}^{s(\epsilon)}_{m},$$

where $H\binom{\ell}{L}{}^{s}_{m}$ is a tail of hypergeometric distribution,
$s(\epsilon) = \left\lfloor \frac{\ell}{L}(m - \epsilon k) \right\rfloor$, $m = L\nu_S(X^L)$.

Remark. Here $\hat{T}(U) = T(U) = \nu_S(U)$.

Example: transductive form of the Law of Large Numbers

Def. The *frequency* of $S \subset \mathbb{X}$ in a finite sample $U \subset \mathbb{X}$:

$$\nu_S(U) = \frac{1}{|U|} \sum_{u \in U} [u \in S].$$

Theorem (common knowledge)

The frequency $\nu_S(X_n^\ell)$ predicts the frequency $\nu_S(X_n^k)$. Prediction accuracy is given by an exact bound

$$\mathsf{P}_n\big[\nu_S(X_n^k) - \nu_S(X_n^\ell) \geq \epsilon\big] = H\big(\tfrac{\ell}{L}\tfrac{s(\epsilon)}{m}\big),$$

where $H\big(\tfrac{\ell}{L}\tfrac{s}{m}\big)$ is a tail of hypergeometric distribution, $s(\epsilon) = \big\lfloor \frac{\ell}{L}(m - \epsilon k) \big\rfloor$, $m = L\nu_S(X^L)$.

Remark. Here $\hat{T}(U) = T(U) = \nu_S(U)$.

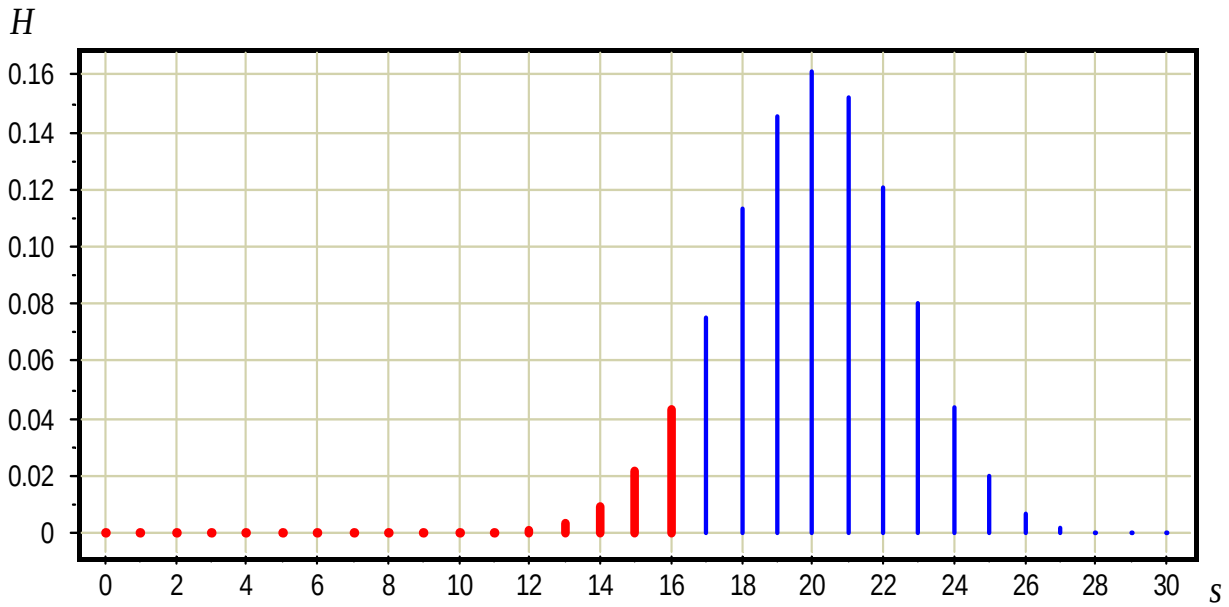
Let us consider the Law of Large Numbers as an example. It is commonly known that the frequency of an event S in the hidden sample [see...] can be predicted by its frequency in the observed sample [see...] .
In classical Probability Theory one predict not frequency but a probability of the event S and one use inequalities of Hoeffding, Bernstain and Chernoff to give asymptotical bounds of the prediction accuracy.

In our transductive framework the bound is given by hypergeometric distribution [see...] . Remarkable than it is an exact [see...] non-asymptotic bound.

Example: transductive form of the Law of Large Numbers

The tail of hypergeometric distribution:

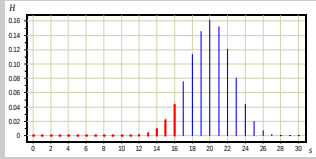
$$H\binom{\ell\ s(\epsilon)}{L\ m} = \sum_{t=s_0}^{s(\epsilon)} \frac{C_m^t C_{L-m}^{\ell-t}}{C_L^\ell}, \quad s_0 = \max\{0, m-k\}$$



Here $\epsilon = 0.05$, $L = 300$, $\ell = 200$, $m = 30$.

The tail of hypergeometric distribution:

$H\binom{\ell}{L}_{m}^{s(\epsilon)} = \sum_{t=s_0}^{s(\epsilon)} \frac{C_m^t C_{L-m}^{\ell-t}}{C_L^\ell}, \quad s_0 = \max\{0, m-k\}$



Here $\epsilon = 0.05$, $L = 300$, $\ell = 200$, $m = 30$.

In this case convergence is a trivial consequence of the fact that the relative width [see...] of hypergeometric peak tends to zero when the sample size L tends to infinity.

Links with classical Kolmogorov's Probability Axioms

Theorem (Correspondence principle)

If we have obtained a bound under Weak PA

$$Q_\epsilon(X^L) = \mathsf{P}_n\big[d(\hat{T}(X_n^\ell), T(X_n^\ell, X_n^k)) > \epsilon\big] \leq \eta(\epsilon, X^L)$$

then analogous bound will be true under Strong PA (classical Kolmogorov's PA)

$$\mathsf{E}_{X^L} Q_\epsilon(X^L) = \mathsf{P}_{X^L}\big[d(\hat{T}(X^\ell), T(X^\ell, X^k)) > \epsilon\big] \leq \mathsf{E}_{X^L}\eta(\epsilon, X^L)$$

Theorem (Correspondence principle)

If we have obtained a bound under Weak PA

$$Q_\epsilon(X^L) = \mathsf{P}_n\left[d(\hat{T}(X_n^\ell), T(X_n^\ell, X_n^k)) > \epsilon\right] \le \eta(\epsilon, X^L)$$

then analogous bound will be true under Strong PA (classical Kolmogorov's PA)

$$\mathsf{E}_{X^L} Q_\epsilon(X^L) = \mathsf{P}_{X^L}\left[d(\hat{T}(X^\ell), T(X^\ell, X^k)) > \epsilon\right] \le \mathsf{E}_{X^L}\eta(\epsilon, X^L)$$

Each bound obtained under Weak Axioms can be easily restated under classical Kolmogorov's Axioms. To do this one may take expectation [see...] of both sides of the inequality.

Links with classical Kolmogorov's Probability Axioms

Theorem (Correspondence principle)

If we have obtained a bound under Weak PA

$$Q_\epsilon(X^L) = \mathsf{P}_n\big[d(\hat{T}(X_n^\ell), T(X_n^\ell, X_n^k)) > \epsilon\big] \leq \eta(\epsilon, X^L)$$

then analogous bound will be true under Strong PA (classical Kolmogorov's PA)

$$\mathsf{E}_{X^L} Q_\epsilon(X^L) = \mathsf{P}_{X^L}\big[d(\hat{T}(X^\ell), T(X^\ell, X^k)) > \epsilon\big] \leq \mathsf{E}_{X^L}\eta(\epsilon, X^L)$$

When a transduction becomes an induction

If $\eta(\epsilon, X^L) \equiv \eta(\epsilon)$ then the same bound is true for any sample.

Theorem (Correspondence principle)

If we have obtained a bound under Weak PA

$$Q_\epsilon(X^L) = \mathsf{P}_n\left[d(\hat{T}(X_n^\ell), T(X_n^\ell, X_n^k)) > \epsilon\right] \le \eta(\epsilon, X^L)$$

then analogous bound will be true under Strong PA (classical Kolmogorov's PA)

$$\mathsf{E}_{X^L} Q_\epsilon(X^L) = \mathsf{P}_{X^L}\left[d(\hat{T}(X^\ell), T(X^\ell, X^k)) > \epsilon\right] \le \mathsf{E}_{X^L}\eta(\epsilon, X^L)$$

When a transduction becomes an induction

If $\eta(\epsilon, X^L) \equiv \eta(\epsilon)$ then the same bound is true for any sample.

Philosophic remark. It is commonly to think that transduction is more restrictive that induction. This is not the case when one write a functional in transductive form and obtain its bound that is true for any sample X^L [see...] . Really this means than transduction transforms into induction.

Links with empirical techniques

Cross-Validation

If we did not succeed to obtain a bound theoretically, or if we obtained very loose (overestimated) bound:

$$Q_\epsilon(X^L) = \mathsf{P}_n\big[d(\hat{T}_n, T_n) > \epsilon\big] \leq \boxed{???},$$

then we can measure it empirically:

$$Q_\epsilon(X^L) \approx \frac{1}{|N'|} \sum_{n \in N'} \big[d(\hat{T}_n, T_n) > \epsilon\big]$$

where $N' \subset \{1, \ldots, N\}$ is small enough to compute the sum and big enough to estimate be accurate.

Cross-Validation

If we did not succeed to obtain a bound theoretically, or if we obtained very loose (overestimated) bound:

$$Q_\epsilon(X^L) = \mathsf{P}_n\big[d(\hat{T}_n, T_n) > \epsilon\big] \le \boxed{???},$$

then we can measure it empirically:

$$Q_\epsilon(X^L) \approx \frac{1}{|N'|} \sum_{n \in N'} \big[d(\hat{T}_n, T_n) > \epsilon\big]$$

where $N' \subset \{1, \ldots, N\}$ is small enough to compute the sum and big enough to estimate be accurate.

The keystone advantage of Weak Probability Axioms. It provides the unique starting point for both theoretical and empirical consideration. If we did not succeed to obtain a bound theoretically, we can measure the prediction functional empirically, replacing the average of all partitions [see...] by the average of some partitions [see...] . This leads to the well known empirical technique — Cross Validation.

The main idea of the further part of presentation is that Cross Validation can help to understand the causes of bounds looseness.

Weak Probability Axioms: pro and con

+ Weak PA is suitable for data analysis, statistics, COLT

Intermediate summary. Our framework is suitable for data analysis, statistics and learning theory, where all samples are finite and all variables can be calculated from data. In classical Probability Theory one operate with hypothetical asymptotic values such that probabilities, expectations, distribution functions, etc. We intend to manage without this hypothetical values. For what reason? Because asymptotic considerations are often the cause of bound looseness and can lead to numerous misunderstandings that are very difficult to reveal.

Weak Probability Axioms: pro and con

+ Weak PA is suitable for data analysis, statistics, COLT
− . . . but not suitable for physics

The same time I think that Weak PA will not suitable for physics and other applications where continuity is crucial. . . so every theory may have its own limitations.

Weak Probability Axioms: pro and con

+ Weak PA is suitable for data analysis, statistics, COLT
− . . . but not suitable for physics
+ Weak PA gives non-asymptotic, exact bounds
− . . . represented by complicated combinatorial formula

Weak Probability Axioms: pro and con

+ Weak PA is suitable for data analysis, statistics, COLT
− . . . but not suitable for physics
+ Weak PA gives non-asymptotic, exact bounds
− . . . represented by complicated combinatorial formula

It can give exact bounds but we are to elaborate fast algorithms to calculate them effectively.

Weak Probability Axioms: pro and con

+ Weak PA is suitable for data analysis, statistics, COLT
− . . . but not suitable for physics
+ Weak PA gives non-asymptotic, exact bounds
− . . . represented by complicated combinatorial formula
+ Weak PA satisfies a "correspondence principle"
− . . . but not all theorems in classical Probability Theory have analogs in Weak PA

Weak Probability Axioms: pro and con

- \+ Weak PA is suitable for data analysis, statistics, COLT
- – . . . but not suitable for physics
- \+ Weak PA gives non-asymptotic, exact bounds
- – . . . represented by complicated combinatorial formula
- \+ Weak PA satisfies a "correspondence principle"
- – . . . but not all theorems in classical Probability Theory have analogs in Weak PA

Each bound obtained under Weak PA can be restated under classical PA. But many measure-specific theorems can not be transferred to the weak form.

Weak Probability Axioms: pro and con

+ Weak PA is suitable for data analysis, statistics, COLT
− . . . but not suitable for physics
+ Weak PA gives non-asymptotic, exact bounds
− . . . represented by complicated combinatorial formula
+ Weak PA satisfies a "correspondence principle"
− . . . but not all theorems in classical Probability Theory have analogs in Weak PA

Open problem:

What part of mathematical statistics can be restated in Weak PA?

Weak Probability Axioms: pro and con

- \+ Weak PA is suitable for data analysis, statistics, COLT
- – . . . but not suitable for physics
- \+ Weak PA gives non-asymptotic, exact bounds
- – . . . represented by complicated combinatorial formula
- \+ Weak PA satisfies a "correspondence principle"
- – . . . but not all theorems in classical Probability Theory have analogs in Weak PA

Open problem:

What part of mathematical statistics can be restated in Weak PA?

A big open problem is "what part of mathematical statistics can be restated in Weak PA?" My opinion is that a sifficiently big part.

The learning problem (classification, regression, etc.)

- $\mathbb{X}$ — set of objects, $\mathbb{Y}$ — set of outputs
- Binary *loss function* $\mathscr{L}: \mathbb{X} \times \mathbb{Y} \to \{0, 1\}$
  - In classification: $\mathscr{L}(x, y) = \left[y \neq y^*(x)\right]$,
  - In regression: $\mathscr{L}(x, y) = \left[|y - y^*(x)| > \delta\right]$
    where $y^*(x)$ — unknown *target function*
- *Training set* $X^\ell = \{x_i\}_{i=1}^\ell \subset \mathbb{X}$ with known losses $\mathscr{L}(x_i, y)$
- *Learning algorithm*

  $\mu: X^\ell \mapsto f$, where $f: \mathbb{X} \to \mathbb{Y}$ — some function
- *Average error* of a function $f: \mathbb{X} \to \mathbb{Y}$ on a set $U \subset \mathbb{X}$

  $$\nu(f, U) = \frac{1}{|U|} \sum_{u \in U} \mathscr{L}(u, f(u))$$
- *Generalization ability*:

  $\nu(\mu X^\ell, U)$ must be sufficiently small for most $U \in \mathbb{X}^*$

The learning problem (classification, regression, etc.)

- $\mathbb{X}$ — set of objects, $\mathbb{Y}$ — set of outputs
- Binary *loss function* $\mathscr{L}: \mathbb{X} \times \mathbb{Y} \to \{0, 1\}$
  - In classification: $\mathscr{L}(x, y) = [y \neq y^*(x)]$,
  - In regression: $\mathscr{L}(x, y) = [|y - y^*(x)| > \delta]$
  where $y^*(x)$ — unknown *target function*
- *Training set* $X^\ell = \{x_i\}_{i=1}^\ell \subset \mathbb{X}$ with known losses $\mathscr{L}(x_i, y)$
- *Learning algorithm*
  $\mu: X^\ell \mapsto f$, where $f: \mathbb{X} \to \mathbb{Y}$ — some function
- *Average error* of a function $f: \mathbb{X} \to \mathbb{Y}$ on a set $U \subset \mathbb{X}$
  $\nu(f, U) = \frac{1}{|U|} \sum_{u \in U} \mathscr{L}(u, f(u))$
- *Generalization ability*:
  $\nu(\mu X^\ell, U)$ must be sufficiently small for most $U \in \mathbb{X}^*$

Now I pass to the second part of my presentation — the Learning Problem. Given a training set X^ℓ [see...] one must learn a function f [see...] that approximates the unknown target function $y^*(x)$ [see...] as well as possible. The approximation quality on a finite sample U is measured by the average error $\nu(f, U)$ [see...] also called the frequency of errors.

The most challenging problem in Learning Theory — how to guarantee that the learned function f will have a small frequency of errors [see...] out of the training set.

Classical bounds

- Vapnik and Chervonenkis (1974):

$$P_\epsilon(F) = \mathsf{P}_{X^L}\Big[\sup_{f\in F}\big(\nu(f,X^k)-\nu(f,X^\ell)\big) > \epsilon\Big]$$
$$\leq \Delta^F(L)\cdot 1.5\, e^{-\epsilon^2\ell};\quad (\text{if } \ell = k)$$

where F is the entire functions set (search space);
$\Delta^F(L)$ — *Global Shatter Coefficient* of the set F,
the number of functions f from F distinguishable on X^L;
$\Delta^F(L) \leq 1.5\,\frac{L^h}{h!}$, $h =$ VC dimension (growth function) of F.

Classical bounds

- Vapnik and Chervonenkis (1974):

$$P_\ell(F) = \mathsf{P}_{X^L}\Big[\sup_{f\in F}\big(\nu(f,X^k)-\nu(f,X^\ell)\big)>\epsilon\Big]$$
$$\le \Delta^F(L)\cdot 1.5\, e^{-\epsilon^2\ell};\ \ (\text{if } \ell = k)$$

where F is the entire functions set (search space);
$\Delta^F(L)$ — *Global Shatter Coefficient* of the set F,
the number of functions f from F distinguishable on X^L;
$\Delta^F(L) \le 1.5\,\frac{L^h}{h!}$, $h =$ VC dimension (growth function) of F.

The classical generalization bound is a product of two terms: complexity term called *shatter coefficient* [see. . .] and convergence term [see. . .] that tends to zero when the sample length tends to infinity.
Let us remind that shatter coefficient of the functions set A is defined as the maximal number of functions from A pairwise indistinguishable on a set X^L.
If the set X^L is fixed then the shatter coefficient is called *local*.
If the set X^L is arbitrary then the shatter coefficient depends on sample size L only (and not on concrete objects) and is called *global* [see. . .] .
In classic VC theory only global shatter coefficient was used.

Classical bounds

- Vapnik and Chervonenkis (1974):

$$P_\epsilon(F) = \mathsf{P}_{X^L}\Big[\sup_{f\in F}\big(\nu(f,X^k)-\nu(f,X^\ell)\big) > \epsilon\Big]$$
$$\leq \Delta^F(L)\cdot 1.5\, e^{-\epsilon^2\ell};\quad (\text{if } \ell = k)$$

where F is the entire functions set (search space);
$\Delta^F(L)$ — *Global Shatter Coefficient* of the set F,
the number of functions f from F distinguishable on X^L;
$\Delta^F(L) \leq 1.5\,\frac{L^h}{h!}$, $h =$ VC dimension (growth function) of F.

- The bound is extremely loose, as it doesn't take into account:
  - the distribution of objects X^ℓ;
  - the target $y^*(x)$;
  - the learning algorithm μ;
  - $1.5\, e^{-\epsilon^2\ell}$ is an asymptotic approximation.

Classical bounds

- Vapnik and Chervonenkis (1974):

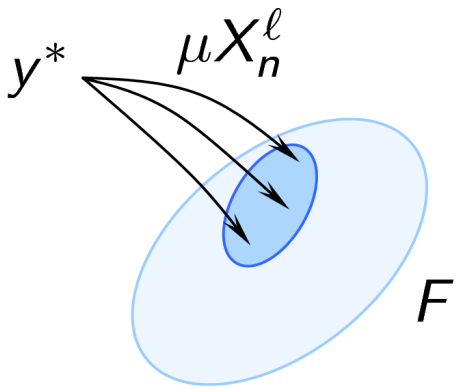
$$P_\epsilon(F) = \mathsf{P}_{X^L}\Big[\sup_{f\in F}\big(\nu(f,X^k)-\nu(f,X^\ell)\big)>\epsilon\Big]$$
$$\leq \Delta^F(L)\cdot 1.5\, e^{-\epsilon^2\ell};\quad (\text{if } \ell = k)$$

  where F is the entire functions set (search space);
  $\Delta^F(L)$ — *Global Shatter Coefficient* of the set F,
  the number of functions f from F distinguishable on X^L;
  $\Delta^F(L) \leq 1.5\,\frac{L^h}{h!}$, $h =$ VC dimension (growth function) of F.
- The bound is extremely loose, as it doesn't take into account:
  - the distribution of objects X^ℓ;
  - the target $y^*(x)$;
  - the learning algorithm μ;
  - $1.5\,e^{-\epsilon^2\ell}$ is an asymptotic approximation.

The main shortcoming of this bound [see. . .] is that the uniform convergence taken in VC theory as an axiom results in extremely overestimated complexity term [see. . .] . Taking supremum [see. . .] one neglect many important peculiarities [see. . .] of the given task.

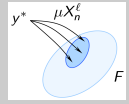
Data-dependent bounds

- *Localization effect:*
  If target y^*, learning algorithm μ,
  and sample X^L are fixed
  then not all functions from F
  can be obtained.



- Uniform convergence bound (Vapnik & Chervonenkis, 1969)
- Theory of learnable (Valiant, 1982)
- First data-dependent bound (?)
- The most tight general VC-like bound (M. Talagrand)
- Self-bounding (Freund)
- Algorithmic luckiness ()
- Computationally tight sample complexity bounds [J. Langford]

- *Localization effect:*
  If target y^*, learning algorithm μ, and sample X^L are fixed then not all functions from F can be obtained.



- Uniform convergence bound (Vapnik & Chervonenkis, 1969)
- Theory of learnable (Valiant, 1982)
- First data-dependent bound (?)
- The most tight general VC-like bound (M. Talagrand)
- Self-bounding (Freund)
- Algorithmic luckiness ()
- Computationally tight sample complexity bounds [J. Langford]

When a particular task is fixed, only a little part of functions [see. . .] can be obtained. This localization effect was understood long ago and several frameworks for data-dependent bound was proposed.

...

Our proposition distinguishes by the total change of Probability Axioms.

Data-dependent bounds under Weak PA

- Weak PA:

$$Q_\epsilon(\mu, X^L) = \mathsf{P}_n\Big[\nu(f_n, X_n^k) - \nu(f_n, X_n^\ell) > \epsilon\Big]$$
$$\leq \Delta_L^\ell(\mu, X^L) \cdot \max_m H\tbinom{\ell}{L}{s(\epsilon)}{m}$$
$$\Big(\leq \Delta^F(L) \cdot 1.5\, e^{-\epsilon^2 \ell}\Big);$$

where $f_n = \mu X_n^\ell$ is a result of learning;
$\Delta_L^\ell(\mu, X^L)$ — *Local Shatter Coefficient* of the set of functions obtainable by learning: $\Big\{f_n \;\Big|\; n = 1, \ldots, N\Big\}$.

Data-dependent bounds under Weak PA

- Weak PA:

$$Q_\epsilon(\mu, X^L) = \mathsf{P}_n\left[\nu(f_n, X_n^k) - \nu(f_n, X_n^\ell) > \epsilon\right]$$
$$\leq \Delta_L^\ell(\mu, X^L) \cdot \max_m H\left(\ell, s(\epsilon)\right)$$
$$\left(\leq \Delta^F(L) \cdot 1.5\, e^{-\epsilon^2 \ell}\right);$$

where $f_n = \mu X_n^\ell$ is a result of learning;
$\Delta_L^\ell(\mu, X^L)$ — *Local Shatter Coefficient* of the set of functions obtainable by learning: $\{f_n \mid n = 1, \ldots, N\}$.

The bound can be obtained under Weak PA [see...] . We will see later that this bound is still very loose. The only advantage of this bound is that this functional [see...] can be measured effectively to understand the causes of bound looseness.

Note that the older data-independent VC bound can be derived from this one.

The next idea is to eliminate maximum on m [see...] in convergence term.

Further tightening data-dependent bound

- **Idea:** the scalar complexity contains too small information about the learning process.
- Splitting the local shatter coeff into *Shatter Profile* $\{D_m\}_{m=0}^{L}$:

$$\Delta_L^\ell(\mu, X^L) = \sum_{m=1}^{L} D_m(\mu, X^L)$$

$D_m(\mu, X^L)$ — local shatter coefficient of the set of functions having m errors on X^L: $\left\{f_n \;\middle|\; \nu(f_n, X^L) = \frac{m}{L},\ n = 1, \ldots, N\right\}$.

Further tightening data-dependent bound

- **Idea:** the scalar complexity contains too small information about the learning process.
- Splitting the local shatter coeff into *Shatter Profile* $\{D_m\}_{m=0}^{L}$:

$$\Delta_L^\ell(\mu, X^L) = \sum_{m=1}^{L} D_m(\mu, X^L)$$

$D_m(\mu, X^L)$ — local shatter coefficient of the set of functions having m errors on X^L: $\{f_n \mid \nu(f_n, X^L) = \frac{m}{L},\ n = 1, \ldots, N\}$.

This idea has yet another interpretation. One scalar characteristic of complexity contains too small information about learning process. It's not sufficient to know, how many different functions can be obtained as a result of learning. Also it is worth to take into account how many functions of a given quality can be obtained.

For this reason we split the local shatter coefficient into $L + 1$ components [see...] . Each component can be considered as a local shatter coefficient of the set of functions that make exactly m errors on the full sample X^L. So we obtain the non-scalar characteristic of complexity that we call *Shatter Profile* [see...] .

Further tightening data-dependent bound

- **Idea:** the scalar complexity contains too small information about the learning process.
- Splitting the local shatter coeff into *Shatter Profile* $\{D_m\}_{m=0}^{L}$:

$$\Delta_L^\ell(\mu, X^L) = \sum_{m=1}^{L} D_m(\mu, X^L)$$

$D_m(\mu, X^L)$ — local shatter coefficient of the set of functions having m errors on X^L: $\left\{f_n \;\middle|\; \nu(f_n, X^L) = \frac{m}{L},\ n = 1, \ldots, N\right\}$.

- Weak PA:

$$Q_\epsilon(\mu, X^L) \leq \sum_{m=1}^{L} D_m(\mu, X^L) \cdot H\binom{\ell\ s(\epsilon)}{L\ m};$$

Further tightening data-dependent bound

- **Idea:** the scalar complexity contains too small information about the learning process.
- Splitting the local shatter coeff into *Shatter Profile* $\{D_m\}_{m=0}^{L}$:

$$\Delta_L^\ell(\mu, X^L) = \sum_{m=1}^{L} D_m(\mu, X^L)$$

$D_m(\mu, X^L)$ — local shatter coefficient of the set of functions having m errors on X^L: $\{f_n \mid \nu(f_n, X^L) = \frac{m}{L},\ n = 1, \ldots, N\}$.

- Weak PA:

$$Q_\epsilon(\mu, X^L) \leq \sum_{m=1}^{L} D_m(\mu, X^L) \cdot H\left(\tfrac{\ell}{L}\tfrac{s(\epsilon)}{m}\right);$$

With shatter profile we obtain a more tight bound [see...] . The previous one could be obtained from this one taken the maximum of convergence term by m [see...] .

The Effective Local Shatter Profile and Coefficient

- **Ideally accurate but "unfair" bound** answers a question: what would be the shatter profile D_m to bound be exact?
- **Theorem**

$$Q_{\epsilon,m}(\mu, X^L) = \mathsf{P}_n\Big[\nu(f_n, X_n^k) - \nu(f_n, X_n^\ell) > \epsilon\Big]\Big[\nu(f_n, X^L) = \frac{m}{L}\Big]$$
$$\leq D_m(\mu, X^L) \cdot H\binom{\ell\ s(\epsilon)}{L\ m};$$

Let us change "$\leq$" by "$=$" and express D_m:

The Effective Local Shatter Profile and Coefficient

- **Ideally accurate but "unfair" bound** answers a question: what would be the shatter profile D_m to bound be exact?
- **Theorem**

$$Q_{\epsilon,m}(\mu, X^L) = \mathsf{P}_n\Big[\nu(f_n, X_n^k) - \nu(f_n, X_n^\ell) > \epsilon\Big]\Big[\nu(f_n, X^L) = \tfrac{m}{L}\Big] \leq D_m(\mu, X^L) \cdot H\big(\tfrac{\ell}{L}\tfrac{s(\epsilon)}{m}\big);$$

Let us change "$\leq$" by "$=$" and express D_m:

To compare proposed bounds empirically we must have an ideally accurate bound as a reference point. Then we introduce a subsidiary functional Q_m [see...] which helps to estimate all components of the Shatter Profile separately [see...] . We call this bound *unfair* because usually one estimate the complexity term to get an upper bound of the quality functional. Whereas here we do a contrary thing: we estimate the quality functional via Cross-Validation [see...] in order to answer a question: what would be the shatter profile D_m [see...] to bound be exact?

The Effective Local Shatter Profile and Coefficient

- **Ideally accurate but "unfair" bound** answers a question: what would be the shatter profile D_m to bound be exact?
- **Theorem**

$$Q_{\epsilon,m}(\mu, X^L) = \mathsf{P}_n\Big[\nu(f_n, X_n^k) - \nu(f_n, X_n^\ell) > \epsilon\Big]\Big[\nu(f_n, X^L) = \tfrac{m}{L}\Big]$$
$$\leq D_m(\mu, X^L) \cdot H\tbinom{\ell\ s(\epsilon)}{L\ m};$$

Let us change "$\leq$" by "$=$" and express D_m:

- *Effective Local Shatter Profile* $\hat{D}_m$, $m = 0, \ldots, L$:

$$\hat{D}_m = \frac{\frac{1}{|N'|}\sum_{n\in N'}\big[\nu(f_n, X_n^k) - \nu(f_n, X_n^\ell) > \epsilon\big]\big[\nu(f_n, X^L) = \frac{m}{L}\big]}{H\tbinom{\ell\ s(\epsilon)}{L\ m}}.$$

- *Effective Local Shatter Coefficient*: $\hat{\Delta}_L^\ell = \hat{D}_0 + \cdots + \hat{D}_L$.

The Effective Local Shatter Profile and Coefficient

- **Ideally accurate but "unfair" bound** answers a question: what would be the shatter profile D_m to bound be exact?
- **Theorem**

$$Q_{\epsilon,m}(\mu, X^L) = \mathsf{P}_n\Big[\nu(f_n, X_n^k) - \nu(f_n, X_n^\ell) > \epsilon\Big]\Big[\nu(f_n, X^L) = \tfrac{m}{L}\Big] \le D_m(\mu, X^L) \cdot H\binom{\ell}{L\,m}^{s(\epsilon)};$$

Let us change "$\le$" by "$=$" and express D_m:

- *Effective Local Shatter Profile* $\hat{D}_m$, $m = 0, \ldots, L$:

$$\hat{D}_m = \frac{\frac{1}{|N'|}\sum_{n\in N'}\big[\nu(f_n, X_n^k) - \nu(f_n, X_n^\ell) > \epsilon\big]\big[\nu(f_n, X^L) = \frac{m}{L}\big]}{H\binom{\ell}{L\,m}^{s(\epsilon)}}.$$

- *Effective Local Shatter Coefficient*: $\hat{\Delta}_L^\ell = \hat{D}_0 + \cdots + \hat{D}_L$.

We call the unfair estimate of shatter profile the *Effective Local Shatter Profile* [see...] . Then we define the *Effective Local Shatter Coefficient* as the sum of all profile components [see...] .

Empirical study

- The goal of experiment — to understand:
  - how tight these bounds are?
  - what effect is more important to bound tightening?
- Bounds to be compared:
  - The worst: classical VC bound
  - Local shatter coefficient bound
  - Local shatter profile bound
  - The best: effective (unfair) local shatter profile bound
- Testing area — the Rule Induction algorithm, because:
  - Global SC is well known
  - Local SC can be easily estimated during rule search
- Testing area — 7 tasks from UCI repository

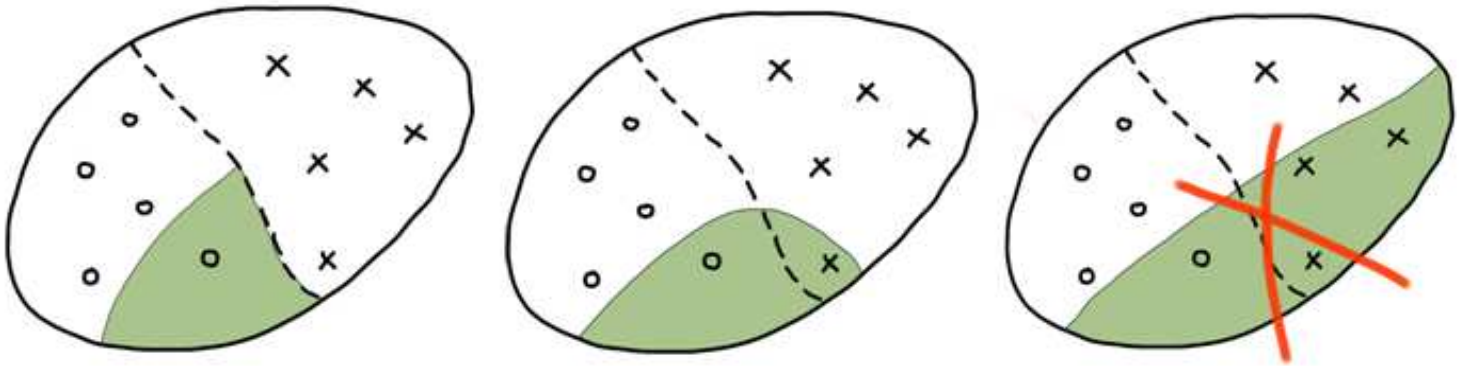
Empirical study
- The goal of experiment — to understand:
  - how tight these bounds are?
  - what effect is more important to bound tightening?
- Bounds to be compared:
  - The worst: classical VC bound
  - Local shatter coefficient bound
  - Local shatter profile bound
  - The best: effective (unfair) local shatter profile bound
- Testing area — the Rule Induction algorithm, because:
  - Global SC is well known
  - Local SC can be easily estimated during rule search
- Testing area — 7 tasks from UCI repository

The goal of our empirical study was to check the looseness of our bounds and to understand, what effects must be taken into account additionally? We compare all three bounds [see...] with an ideal [see...] .

We choose a Rule induction System as a Testing Area because global and local shatter coefficients can be easily obtained for this kind of learning algorithms.

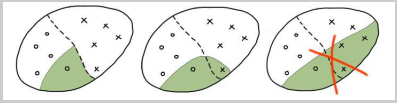
The rule definition

- Features $f_1(x), \ldots, f_n(x)$.
- *Rule* ϕ is a conjunction: $\phi(x, \theta) = \bigwedge_{f \in \Omega} \left[f(x) \leqslant \theta_f \right]$,
  where $\Omega \subset \{f_1, \ldots, f_n\}$, θ_f — threshold for feature $f(x)$.
- Rule $\phi(x, \theta)$ is well interpretable while $|\Omega| \lesssim 5$.
- $\phi(x) = 1 \iff$ Rule $\phi(x)$ *covers* object $x \in \mathbb{X}$.
- *Rule* $\phi_y(x)$ *of class* y covers many objects from y
  and none or a few objects from $\mathbb{Y} \setminus y$:



The rule definition

- Features $f_1(x), \ldots, f_n(x)$.
- *Rule* ϕ is a conjunction: $\phi(x,\theta) = \bigwedge_{f \in \Omega} [f(x) \lessgtr \theta_f]$, where $\Omega \subset \{f_1, \ldots, f_n\}$, θ_f — threshold for feature $f(x)$.
- Rule $\phi(x,\theta)$ is well interpretable while $|\Omega| \lesssim 5$.
- $\phi(x) = 1 \Longleftrightarrow$ Rule $\phi(x)$ *covers* object $x \in \mathbb{X}$.
- *Rule* $\phi_y(x)$ *of class* y covers many objects from y and none or a few objects from $\mathbb{Y} \setminus y$:

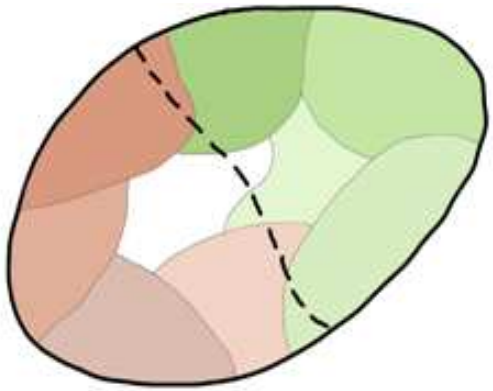


A few words about what is rule induction. A rule is a well interpretable predicate, usually a conjunction, that covers many objects of one class and none [see...] or a few [see...] objects of other classes. For example this [see...] is not a rule because it covers both classes significantly.

Classifier is a combination of rules

- Decision List of rules:

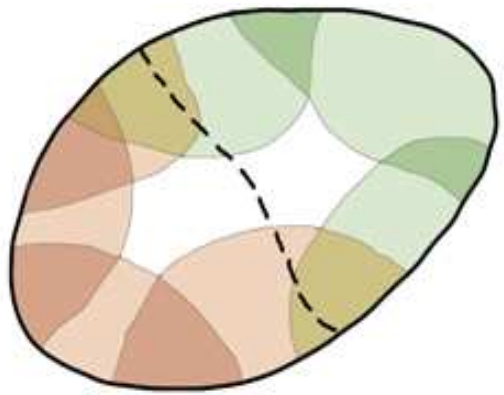
$$\begin{aligned}
&\texttt{If}\ \ \phi^1_{y_1}(x) \to f(x) := y_1\,;\\
&\texttt{If}\ \ \phi^2_{y_2}(x) \to f(x) := y_2\,;\\
&\ldots\\
&\texttt{If}\ \ \phi^T_{y_T}(x) \to f(x) := y_T\,;
\end{aligned}$$



- Weighted Voting of rules:

$$f(x) = \arg\max_{y\in\mathbb{Y}} \sum_{t=1}^{T_y} w_y^t \phi_y^t(x),$$

where
$\phi_y^t(x)$ — t-th rule of class y;
w_y^t — its weight.



Classifier is a combination of rules

- Decision List of rules:

  If $\phi^1_{y_1}(x) \to f(x) := y_1$;
  If $\phi^2_{y_2}(x) \to f(x) := y_2$;
  ...
  If $\phi^T_{y_T}(x) \to f(x) := y_T$;

- Weighted Voting of rules:

  $$f(x) = \arg\max_{y \in Y} \sum_{t=1}^{T_y} w^t_y \phi^t_y(x),$$

  where
  $\phi^t_y(x)$ — t-th rule of class y;
  w^t_y — its weight.

A rule-based classifier can be considered as an ensemble of rules based on the principle of seniority voting [see...] or majority voting [see...] .

Results

Task	L	Glob.SC	Loc.SC	Loc.S.Profile	Eff.Loc.SC
crx	690	$2.8 \cdot 10^8$	$3.5 \cdot 10^4$	$1.1 \cdot 10^4$	21 ± 11
german	1000	$5.2 \cdot 10^8$	$3.1 \cdot 10^4$	$1.8 \cdot 10^4$	47 ± 38
hepatitis	155	$5.5 \cdot 10^6$	$1.8 \cdot 10^4$	$8.4 \cdot 10^3$	58 ± 46
horse-colic	300	$1.9 \cdot 10^6$	$1.3 \cdot 10^4$	$6.3 \cdot 10^3$	5 ± 3
hypothyroid	3163	$5.3 \cdot 10^8$	$2.2 \cdot 10^4$	$9.2 \cdot 10^3$	43 ± 28
liver	345	$1.5 \cdot 10^7$	$2.9 \cdot 10^4$	$9.5 \cdot 10^3$	12 ± 8
promoters	106	$4.4 \cdot 10^9$	$5.3 \cdot 10^4$	$2.4 \cdot 10^4$	13 ± 4

Interpretation

In all tasks effective local SC $\ll L$.
Then the "effective local VC dimension" newer exceeds 1 !

Results

Task	L	Glob.SC	Loc.SC	Loc.S.Profile	Eff.Loc.SC
crx	690	$2.8 \cdot 10^8$	$3.5 \cdot 10^4$	$1.1 \cdot 10^4$	21 ± 11
german	1000	$5.2 \cdot 10^8$	$3.1 \cdot 10^4$	$1.8 \cdot 10^4$	47 ± 38
hepatitis	155	$5.5 \cdot 10^6$	$1.8 \cdot 10^4$	$8.4 \cdot 10^3$	58 ± 46
horse-colic	300	$1.9 \cdot 10^6$	$1.3 \cdot 10^4$	$6.3 \cdot 10^3$	5 ± 3
hypothyroid	3163	$5.3 \cdot 10^8$	$2.2 \cdot 10^4$	$9.2 \cdot 10^3$	43 ± 28
liver	345	$1.5 \cdot 10^7$	$2.9 \cdot 10^4$	$9.5 \cdot 10^3$	12 ± 8
promoters	106	$4.4 \cdot 10^9$	$5.3 \cdot 10^4$	$2.4 \cdot 10^4$	13 ± 4

Interpretation

In all tasks effective local SC $\ll L$.
Then the "effective local VC dimension" newer exceeds 1 !

We measured complexity terms for seven real tasks from UCI repository [see...] and four types of bounds.
The result is choking.
None of the bounds can be called tight!

Moreover, the effective local shatter coefficient [see...] is always significantly less than the sample size [see...] . This means that the effective local VC dimension never exceeds 1. So the VC dimension bears no relation to generalization bounds.

Directions of further investigation

- So, what causes of bound looseness did we miss?
  I have two ideas:

Directions of further investigation

- So, what causes of bound looseness did we miss?

  I have two ideas:
  - Functions from $\{f_n \mid n = 1, \ldots, N\}$ have different chance to be obtained as a result of learning.
  - Two similar functions from $\{f_n \mid n = 1, \ldots, N\}$ really are not 2 but ≈ 1 function.

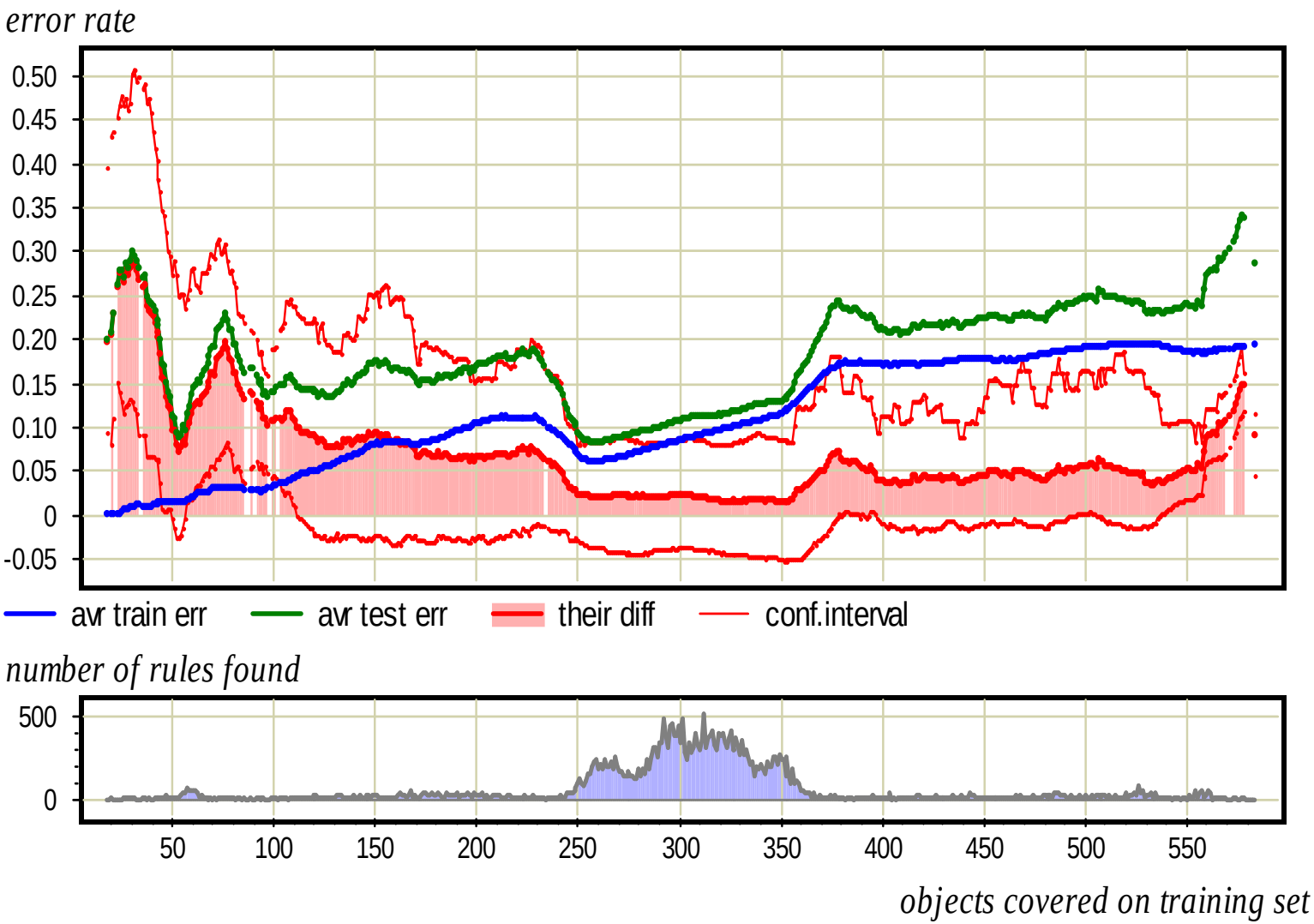
What can we do today

- Define the *overfitting* of a rule $\phi(x)$ as

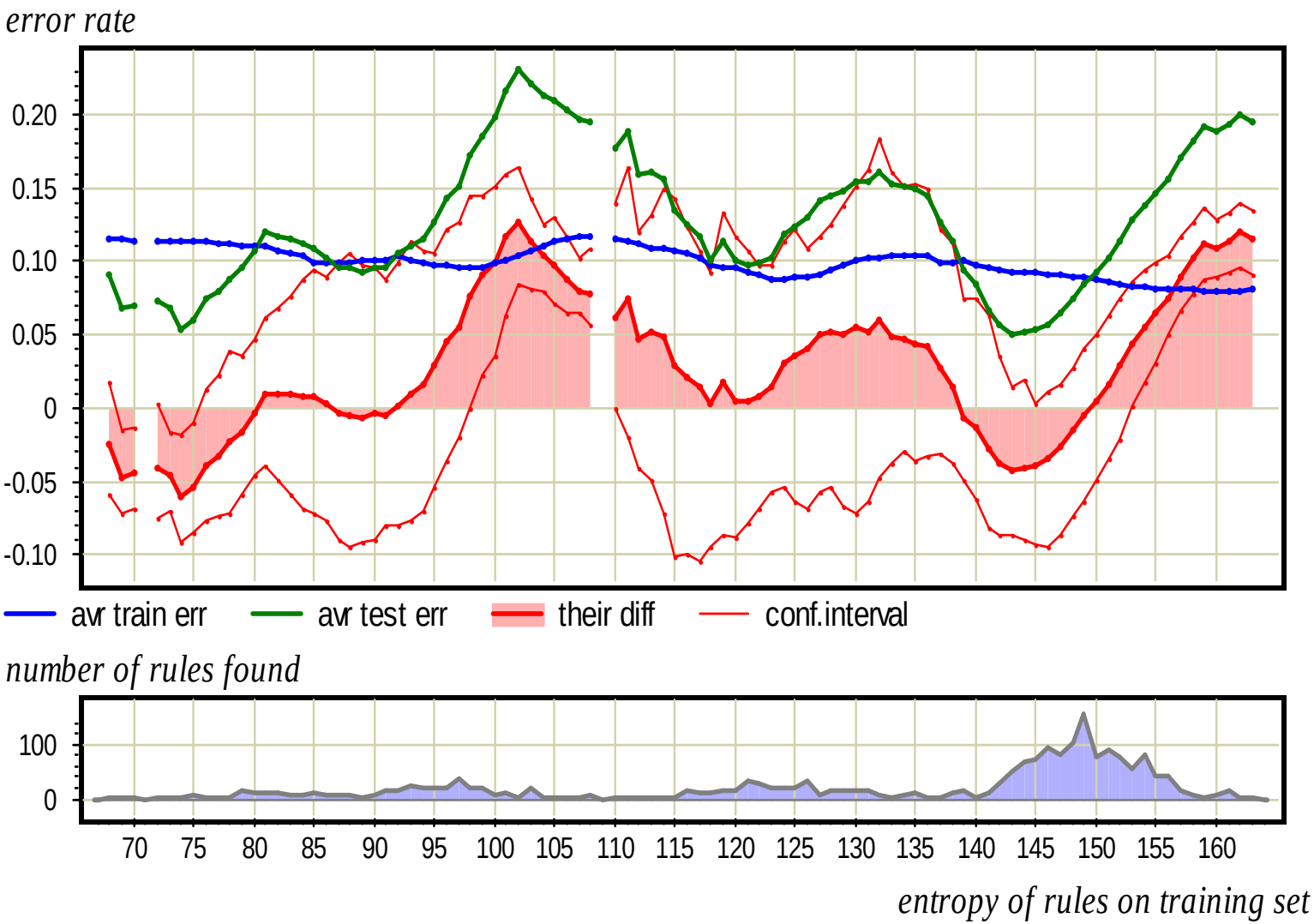
$$\delta(\phi, X^\ell, X^k) = \nu(\phi, X^k) - \nu(\phi, X^\ell)$$

- Estimate (nonparametric) regression:
  how δ depends on rules properties measured on training set X_n^ℓ during inductive search:
  - number of errors made on X_n^ℓ;
  - number of objects covered on X_n^ℓ;
  - number of terms in a rule;
  - entropy of a rule on X_n^ℓ;
  - breadth and width of the inductive search;
- We use these technique for:
  - Controlling rules overfitting
  - Risk Assessment — estimation of PD (probability of default) in Credit Scoring application: $PD(x) = \nu(\phi, X^\ell) + \hat{\delta}(\phi)$

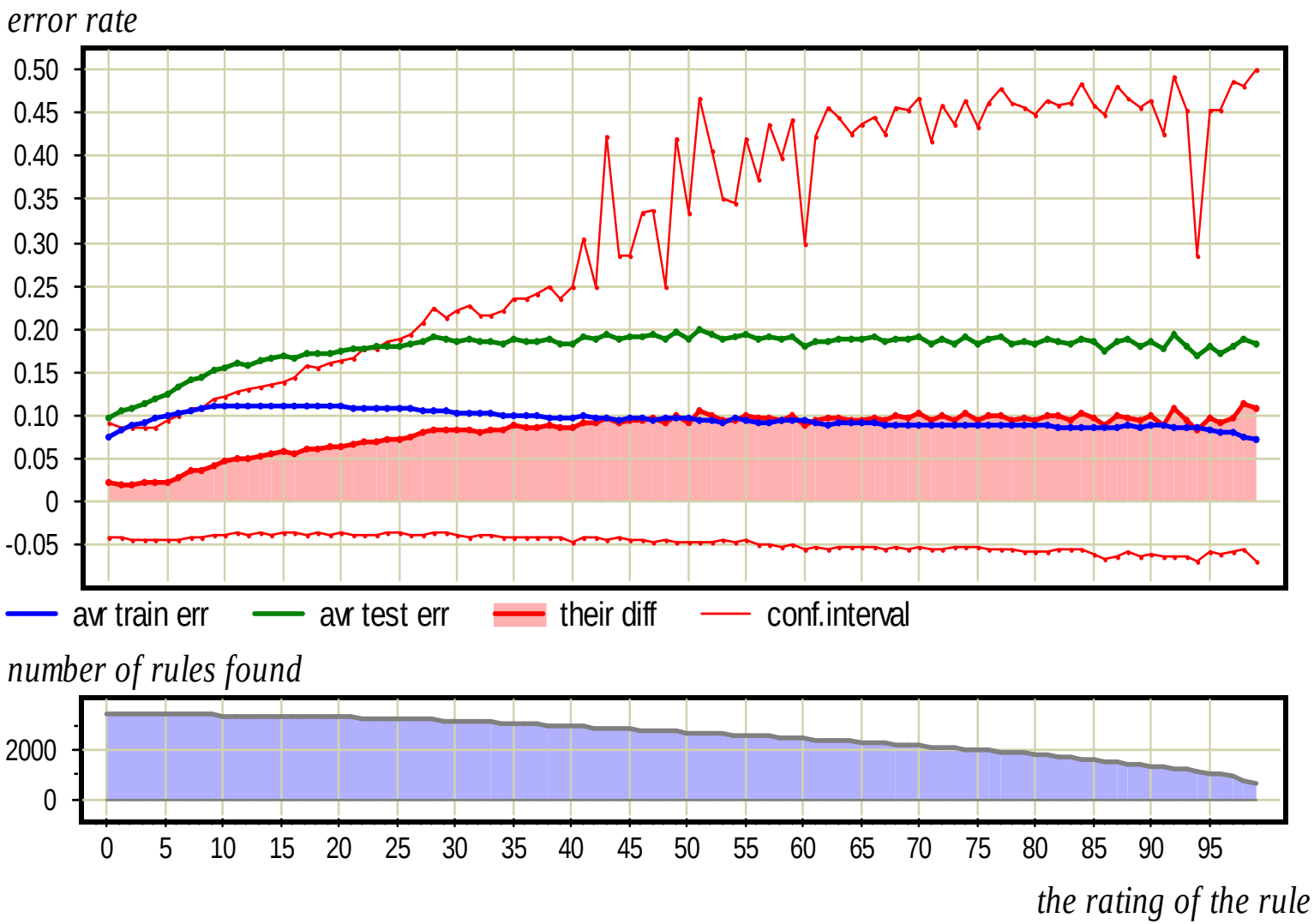
Example 1


error rate
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0
-0.05
avr train err
avr test err
their diff
conf.interval
number of rules found
500
0
50
100
150
200
250
300
350
400
450
500
550
objects covered on training set

Example 2


error rate
0.20
0.15
0.10
0.05
0
-0.05
-0.10
avr train err
avr test err
their diff
conf.interval
number of rules found
100
0
70
75
80
85
90
95
100
105
110
115
120
125
130
135
140
145
150
155
160
entropy of rules on training set

Example 3


error rate
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0
-0.05
avr train err
avr test err
their diff
conf.interval
number of rules found
2000
0
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
the rating of the rule