Valid Predictions for Diagnostic Purposes

Alex Gammerman
(in collaboration with V.Vovk, I.Nouretdinov, V.Fedorova)

Computer Learning Research Centre
and Department of Computer Science
Royal Holloway, University of London

alex@cs.rhul.ac.uk

Frontiers in Diagnostic Technologies
2nd International Conference
November 2011, Frascati

Content

- Introduction
- Conformal Predictors (" Confidence Machine"): basic ideas
- Applications in Diagnostics
  - Data Analysis in Plasma (next talk by Jesus Vega, Andrea Murari et al.)
  - Medical: clinical, MRI, proteomics-based
  - Environment
  - Other Applications: House Market, Household Analysis, PPI and String Kernels, Image Classification, Network Traffic, etc.

Introduction

Problem of prediction - classification and regression.

Classical classification and regression techniques can deal with conventional small-scale, low-dimensional data sets.

To apply to modern **high-dimensional** and high-throughput data sets encounter serious conceptual and computational difficulties.

New techniques: support vector machines (Vapnik, 1995, 1998); other kernel methods to deal with high-dimensional data sets.

Drawback: lack of useful measures of confidence in their predictions.

For example, bounds in PAC theory on the probability of error exceed 1.

On-line prediction protocol:

$\mathrm{Err}_0 := 0$; $\mathrm{Mult}_0 := 0$; $\mathrm{Emp}_0 := 0$;
FOR $n = 1, 2, \ldots$:
Reality outputs $x_n \in \mathbf{X}$;
Predictor outputs $\Gamma^\epsilon_n \subseteq \mathbf{Y}$ for all $\epsilon \in (0, 1)$;
Reality outputs $y_n \in \mathbf{Y}$;

$$\mathrm{err}^\epsilon_n := \begin{cases} 1 & \text{if } y_n \notin \Gamma^\epsilon_n \\ 0 & \text{otherwise,} \end{cases} \qquad \mathrm{Err}^\epsilon_n := \mathrm{Err}^\epsilon_{n-1} + \mathrm{err}^\epsilon_n, \quad \epsilon \in (0, 1);$$

$$\mathrm{mult}^\epsilon_n := \begin{cases} 1 & \text{if } |\Gamma^\epsilon_n| > 1 \\ 0 & \text{otherwise,} \end{cases} \qquad \mathrm{Mult}^\epsilon_n := \mathrm{Mult}^\epsilon_{n-1} + \mathrm{mult}^\epsilon_n, \quad \epsilon \in (0, 1);$$

$$\mathrm{emp}^\epsilon_n := \begin{cases} 1 & \text{if } |\Gamma^\epsilon_n| = 0 \\ 0 & \text{otherwise,} \end{cases} \qquad \mathrm{Emp}^\epsilon_n := \mathrm{Emp}^\epsilon_{n-1} + \mathrm{Emp}^\epsilon_n, \quad \epsilon \in (0, 1).$$

This talk: to "hedge" the predictions—to complement with measures of their accuracy and reliability (**"confidence machines"** or **"conformal predictors"**).

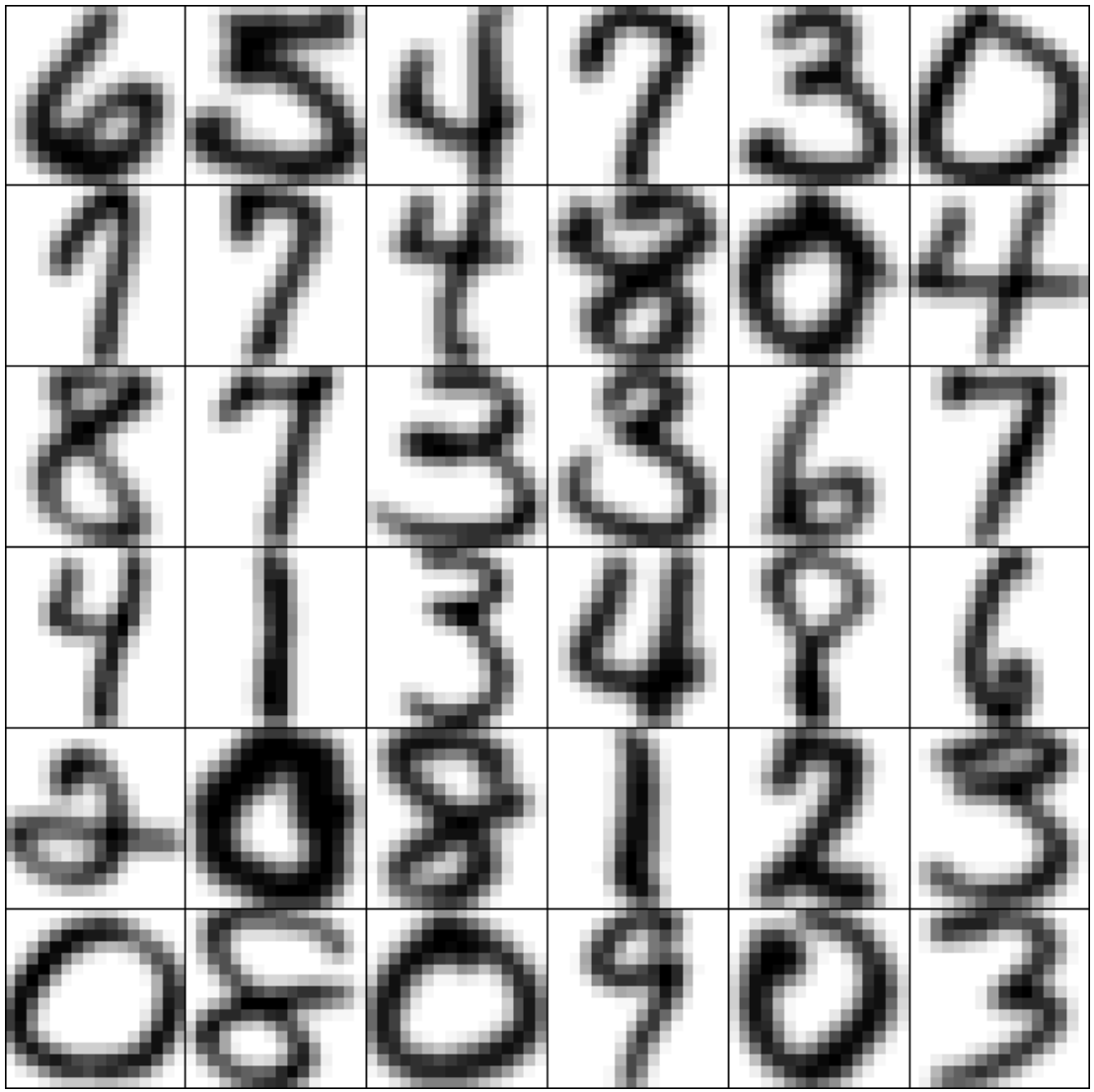
These measures: valid, informative, tailored to the individual object to be predicted.

Their most important property is the automatic **validity** under the randomness assumption (the data are i.i.d.): they never overrate the accuracy and reliability of their predictions.

Another property: **efficiency**.

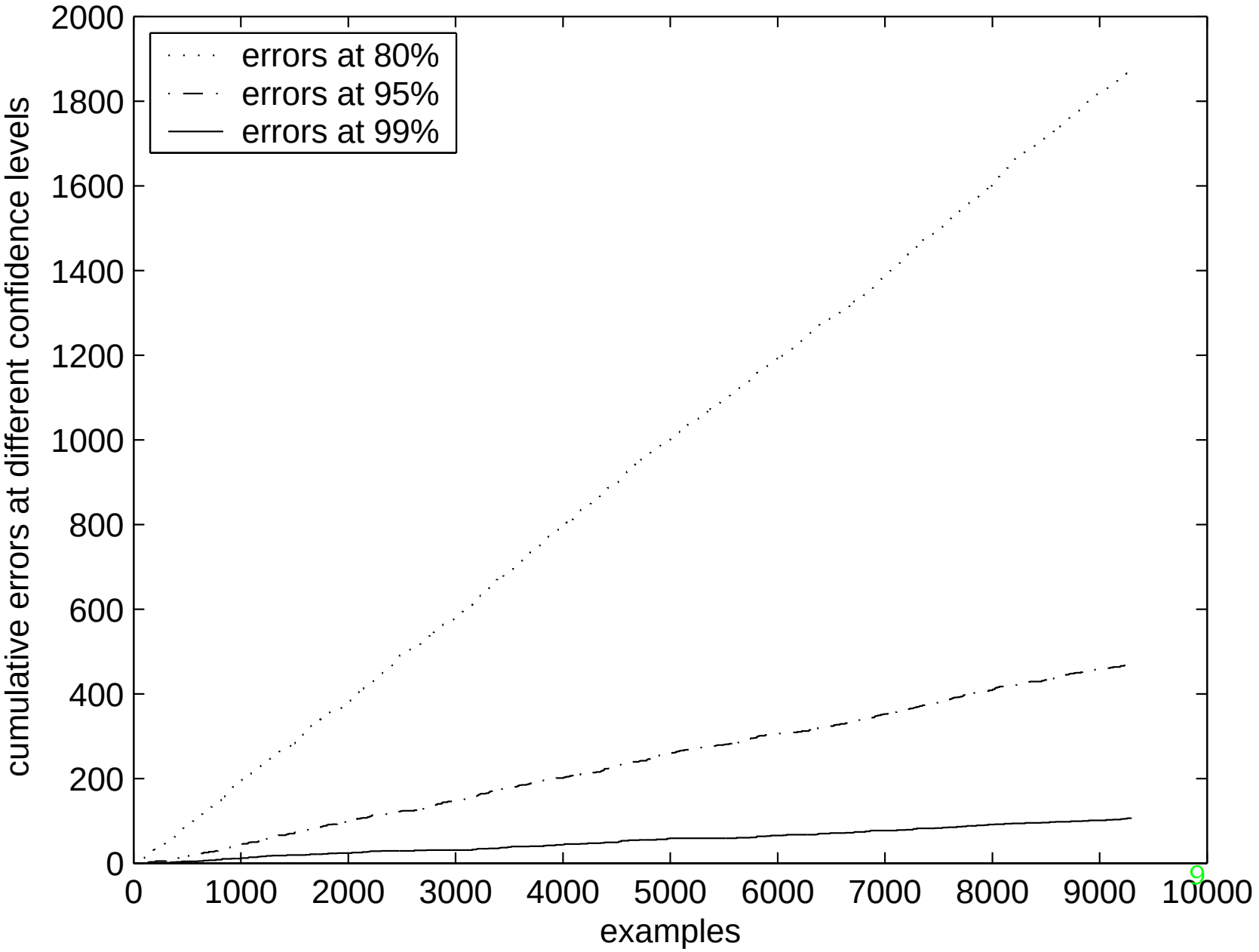
Any classification or regression algorithm can be transformed into a conformal predictor.

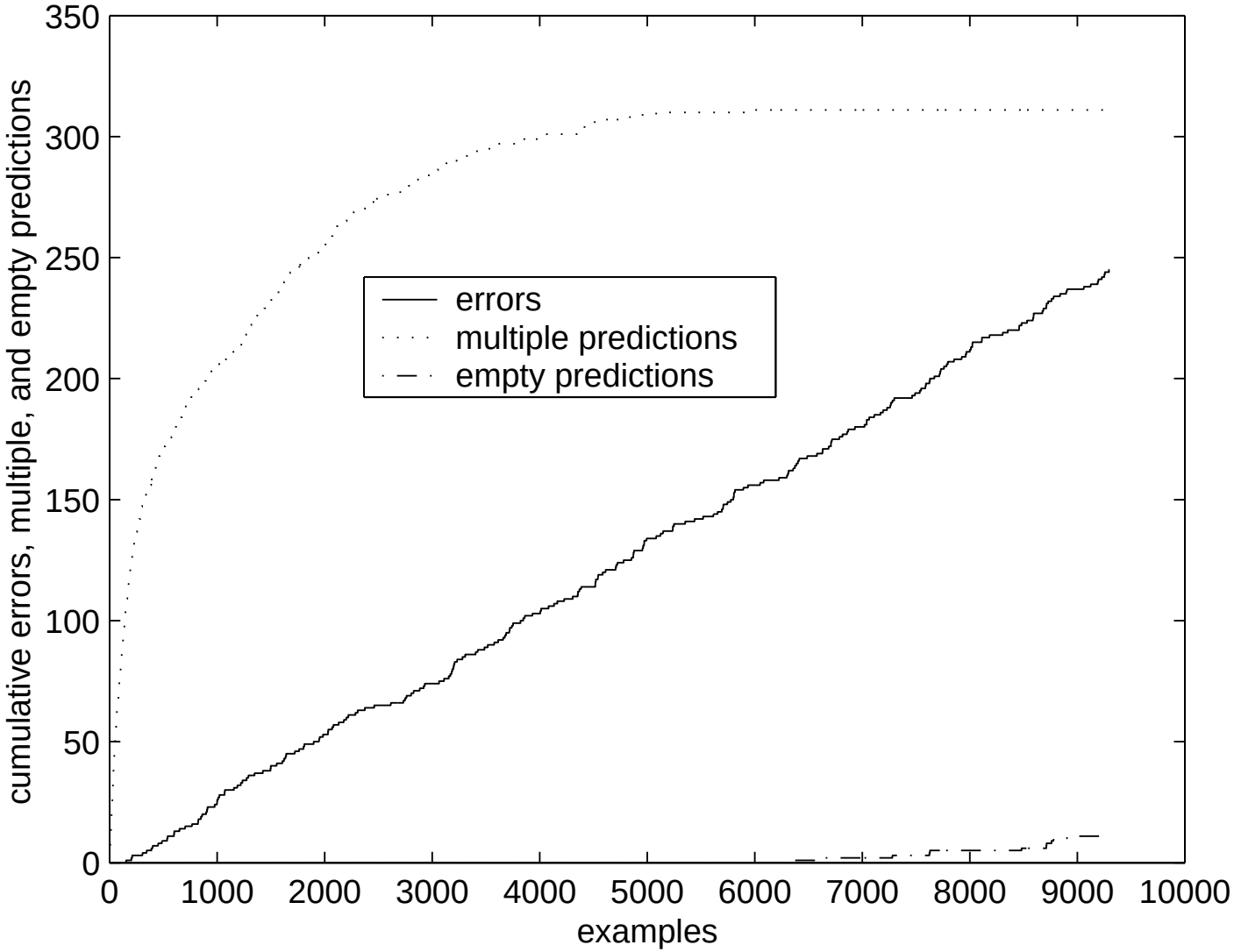
Example: hand-written digits - USPS (9298 digits).



Example: hand-written digits: the cumulative numbers of errors for the 1-nearest neighbour conformal predictor on the USPS data set (9298 hand-written digits, randomly permuted). The solid line for 99%, the dash-dot line for 95%, and the dotted line for 80%.

Next two slides: **validity** and **efficiency**


errors at 80%
errors at 95%
errors at 99%
cumulative errors at different confidence levels
examples
0
200
400
600
800
1000
1200
1400
1600
1800
2000
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000


cumulative errors, multiple, and empty predictions
examples
errors
multiple predictions
empty predictions
0
50
100
150
200
250
300
350
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000

Classification. The idea is to try every possible label Y as a candidate for x_{l+1}'s label and see how well the resulting sequence

$$(x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, Y)$$

conforms to the randomness assumption (if it does conform to this assumption, we will say that it is "random").

The ideal case is where all Ys but one lead to sequences that are not random.

We can then use the remaining Y as a confident prediction for y_{l+1}.

Problem of hedged prediction $\rightarrow$ the problem of testing randomness.

"Universal" notion of randomness by Kolmogorov, Martin-Löf and Levin based on the existence of universal Turing machines.

Let $\mathbf{Z}$ be the set of all possible examples; as each example consists of an object and a label, $\mathbf{Z} = \mathbf{X} \times \mathbf{Y}$, where $\mathbf{X}$ is the set of all possible objects and $\mathbf{Y}$, $|\mathbf{Y}| > 1$, is the set of all possible labels.

Martin-Löf's definition

A function $t : \mathbf{Z}^* \to [0, 1]$ is a randomness test if

1. for all $\epsilon \in (0, 1)$, all $n \in \{1, 2, \dots\}$ and all probability distributions P on $\mathbf{Z}$,

$$P^n \{z \in \mathbf{Z}^n : t(z) \leq \epsilon\} \leq \epsilon; \tag{1}$$

2. t is upper semicomputable.

Prediction with confidence and credibility

- consider all possible values $Y \in \mathbf{Y}$ for the label y_{l+1};
- find the randomness level detected by t for every possible completion $(x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, Y)$;
- predict the label Y corresponding to a completion with the largest randomness level detected by t;
- output as the confidence in this prediction one minus the second largest randomness level detected by t;
- output as the credibility of this prediction the randomness level detected by t of the output prediction Y (i.e., the largest randomness level detected by t over all possible labels).

Selected test examples from the USPS data set

The p-values of digits (0–9), label (true / predicted), confidence and credibility:

0	1	2	3	4	5	6	7	8	9	label	conf.	cr
0.0%	0.1%	0.0%	0.0%	0.1%	0.0%	100%	0.0%	0.0%	0.0%	6 / 6	99.9%	10
0.3%	0.4%	1.1%	0.7%	1.4%	0.7%	0.4%	0.3%	0.7%	0.8%	6 / 4	98.9%	1.4
0.0%	0.3%	0.0%	0.0%	0.2%	0.0%	0.0%	0.0%	0.1%	100%	9 / 9	99.7%	10

Computed using the support vector method with the polynomial kernel of degree 5.

Conformal prediction from support vector machines

$$\frac{1}{2}(w \cdot w) + C\left(\sum_{i=1}^{l} \xi_i\right) \to \min$$

(where C is a fixed constant) subject to constraints

$$y_i\left((x_i \cdot w) + b\right) \geq 1 - \xi_i,\ i = 1, \ldots, l.$$

With each data set

$$(x_1, y_1), \ldots, (x_n, y_n)$$

one associates an optimization problem whose solution produces nonnegative numbers $\alpha_1, \ldots, \alpha_n$ ("Lagrange multipliers").

Taking the completion

$$(x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, Y)$$

as our data set, we can find the corresponding $\alpha_1, \ldots, \alpha_{l+1}$.

If Y is different from the actual label y_{l+1}, we expect (x_{l+1}, Y) to be an outlier in the completion and so α_{l+1} be large as compared with $\alpha_1, \ldots, \alpha_l$.

A natural way to compare α_{l+1} to the other αs is to look at the ratio

$$p_Y := \frac{\left|\{i = 1, \ldots, l+1 : \alpha_i \geq \alpha_{l+1}\}\right|}{l+1},$$

which we call the p-value associated with the possible label Y for x_{l+1}. In words, the p-value is the proportion of the αs which are at least as large as the last α.

General definition

A nonconformity measure is a function that assigns to every data sequence a sequence of numbers $\alpha_1, \ldots, \alpha_n$, called nonconformity scores, such that:

interchange of any two examples (x_i, y_i) and (x_j, y_j) leads to the interchange of the corresponding nonconformity scores (with all the other nonconformity scores unchanged).

The corresponding conformal predictor maps each data set

$$(x_1, y_1), \ldots, (x_l, y_l),$$

$l = 0, 1, \ldots$, each new example x_{l+1}, and each confidence level $1 - \epsilon \in (0, 1)$, into the prediction set

$$\Gamma^\epsilon \left(x_1, y_1, \ldots, x_l, y_l, x_{l+1} \right) := \{ Y \in \mathbf{Y} : p_Y > \epsilon \},$$

where p_Y are defined by

$$p_Y := \frac{\left| \{ i = 1, \ldots, l+1 : \alpha_i \geq \alpha_{l+1} \} \right|}{l+1}$$

with $\alpha_1, \ldots, \alpha_{l+1}$ being the nonconformity scores corresponding to the completion

$$(x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, Y).$$

Associating with each completion its p-value gives a randomness test.

Therefore: for each l the probability of the event

$$y_{l+1} \in \Gamma^\epsilon\left(x_1, y_1, \ldots, x_l, y_l, x_{l+1}\right)$$

is at least $1 - \epsilon$.

In the case of classification we can summarize the prediction sets Γ^ϵ by two numbers: the confidence

$$\sup\left\{1 - \epsilon : |\Gamma^\epsilon| \leq 1\right\}$$

and the credibility

$$\inf\left\{\epsilon : |\Gamma^\epsilon| = 0\right\}.$$

Efficient Regression

We cannot consider all possible values Y for y_{l+1}: infinitely many of them. However, there might still be efficient ways to compute the prediction sets Γ^ϵ.

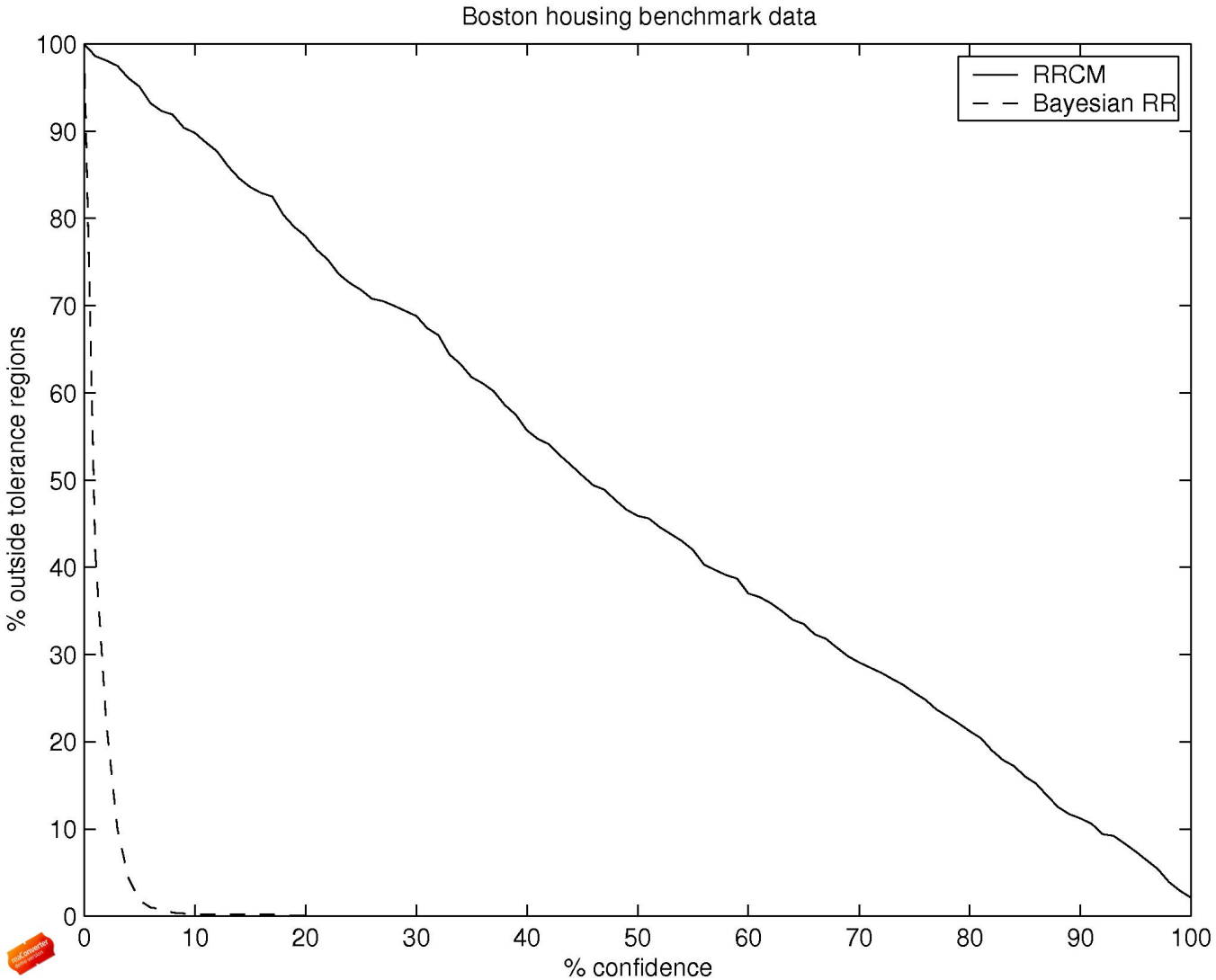
The idea is that if α_i are defined as the residuals

$$\alpha_i = |y_i - f_Y(x_i)| \tag{2}$$

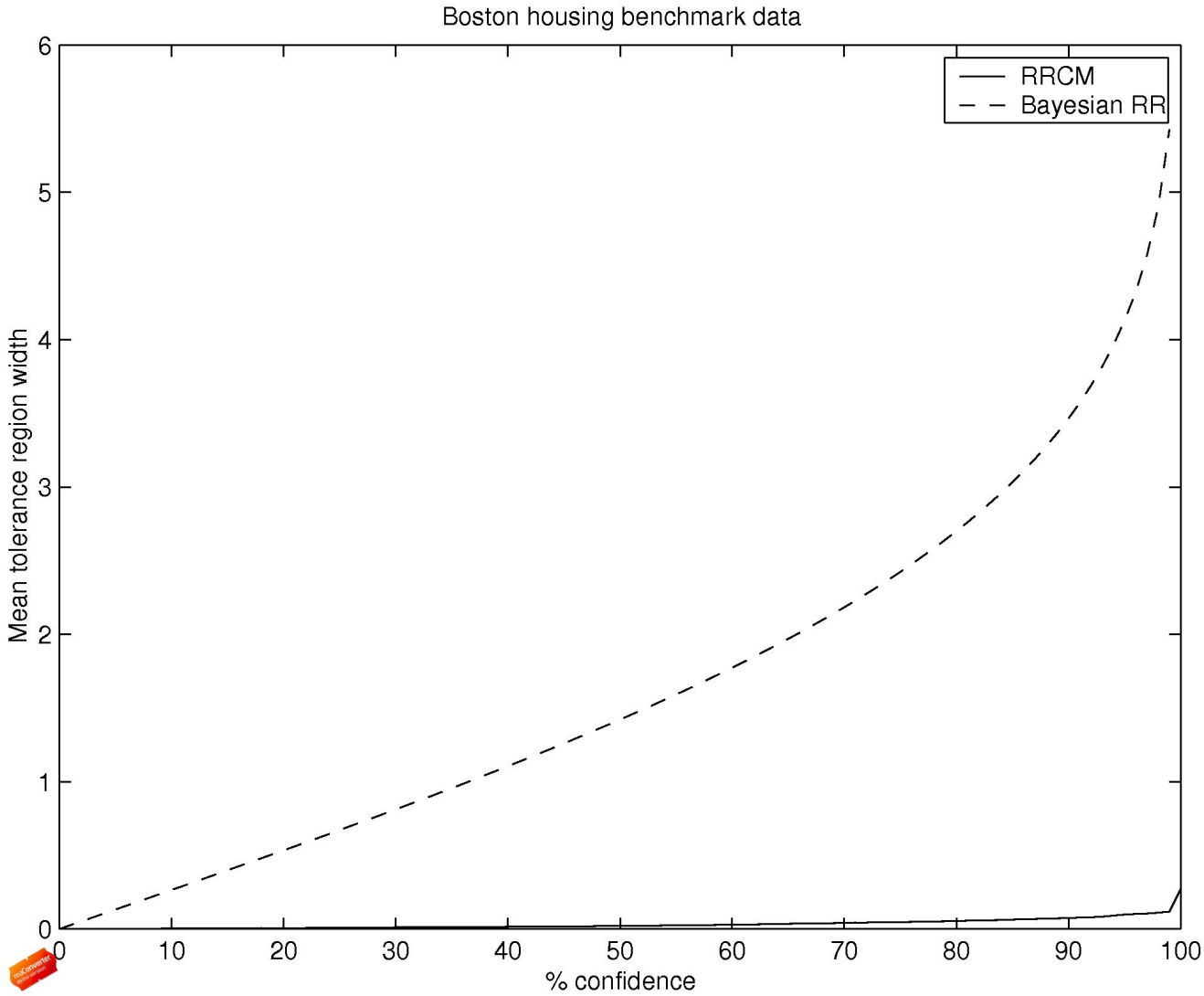
where $f_Y : \mathbf{X} \to \mathbb{R}$ is a regression function then α_i may have a simple expression in terms of Y, leading to an efficient way of computing the prediction sets.

Implemented in in the case where f_Y is found from the ridge regression, or kernel ridge regression with the resulting algorithm of hedged prediction called the *ridge regression confidence machine*.

RRCM validity



RRCM efficiency



APPLICATIONS

- Data Analysis in Plasma (next talk by Jesus Vega)
- Image Classification
- Medical: clinical, MRI, proteomics-based
- Newtwork Traffic

- Environment
- House Market
- Household Analysis
- Biology: PPI and String Kernels
- Other Applications

Clinical Diagnostic of Acute Abdominal Pain based on Conformal Predictors

Prediction for an individual patient:

APP	DIV	PPU	NAP	CHO	INO	PAN	RCO	DYS	true label
1.2%	0.4%	0.2%	2.8%	5.7%	0.9%	1.4%	0.5%	80.6%	DYS

- At the confidence level 95% the prediction region is multiple, {cholecystitis, dyspepsia}.
- When we relax the confidence level to 90%, the prediction region narrows down to {dyspepsia};
- at the confidence level 99% the prediction region widens to {appendicitis, non-specific abdominal pain, cholecystitis, pancreatitis, dyspepsia}.

Different presentation: the patient has DYS with conf=94.3% and cred=80.6%.

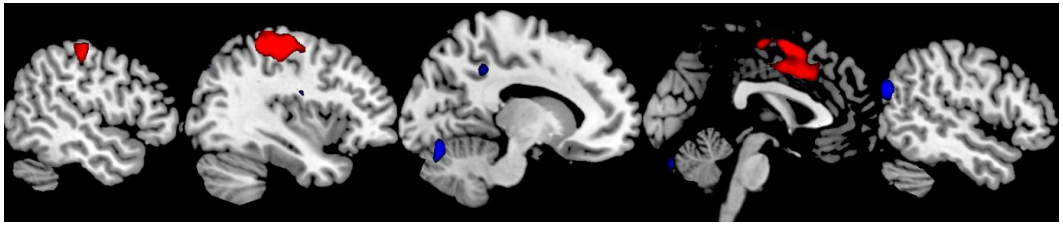
Image Diagnostic based on Conformal Predictors

fMRI-based diagnostic of depression

Ilia Nouretdinov, Sergi G. Costafreda, Alexander Gammerman, Alexey Chervonenkis, Vladimir Vovk, Vladimir Vapnik, and Cynthia H.Y. Fu.
Machine learning classification with confidence: Application of transductive conformal predictors to MRI-based diagnostic and prognostic markers in depression.
NeuroImage, 56(2):809 813, 2011.

MRI-based diagnostic

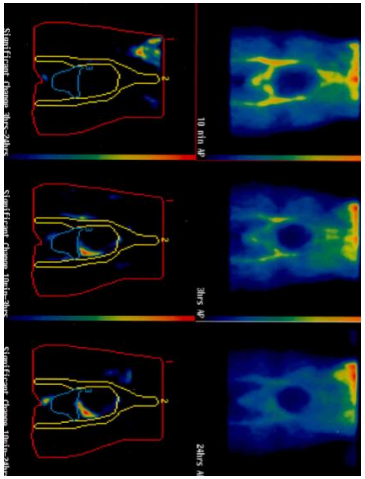
Diagnostic classification of depression from functional MRI.



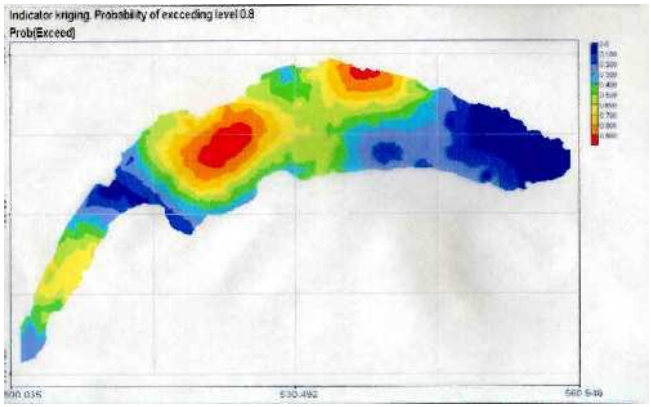
fMRI BOLD responses during implicit processing of sad facial expressions of high intensity. Increased BOLD (blood oxygenation level dependent) in left superior frontal gyrus (in red) and other parts of brain were predictive of a diagnosis of depression.

Proteomics of OC

Ovarian Cancer – CP based on Proteomics Data



Lake Geneva


Indicator kriging. Probability of exceeding level 0.8
Prob[Exceed]

Classification for tj-ii thomson scattering

J. Vega, A. Murari, A. Pereira, S. Gonzalez, and I. Pastor. **Accurate and reliable image classification by using conformal predictors in the tj-ii thomson scattering**. Review of Scientific Instruments, 81(10):10E118 10E1184, 2010.

Conclusion

Many machine-learning techniques can be complemented with provably valid measures of accuracy and reliability.

This talk: Support Vector Machines, and the ridge regression procedure; but the principle is general: virtually any successful prediction technique designed to work under the randomness assumption can be used to produced equally successful hedged predictions.

Replacing the original simple predictions with hedged predictions enables us to control the number of errors made by appropriately choosing the confidence level.

Current: Testing Randomness and other assumptions: martingales.

References

- Alexander Gammerman and Vladimir Vovk. "Hedging Predictions in Machine Learning", The Computer Journal (2007) 50 (2): 151-163.

- V. Vovk, A.Gammerman, G.Shafer (2005) Algorithmic learning in a random world. New York: Springer