Discretization of continuous attributes for supervised learning
Variance evaluation and variance reduction

Louis WEHENKEL - University of Liège - Belgium

- Motivation
- Experimental setup
- Crisp discretization and its variance
- Fuzzy discretization and its variance
- Variance reduction of crisp thresholds
- Bayesian discretization

Motivation

- Symbolic machine learning : main feature is interpretability of results
- Results : parameters used to formulate rules
  - Attributes and thresholds selected to formulate rules’ conditions
  - Probabilities or set membership degrees attached to rules’ conclusions
- Results depend (often too strongly) on the learning sample used

⇒ Results are not as interpretable as expected

- Questions :
  - How much do they depend ⇒ parameter variance ?
  - Is it possible to reduce parameter variance without losing in accuracy ?

⇒ Experimental study of threshold variance

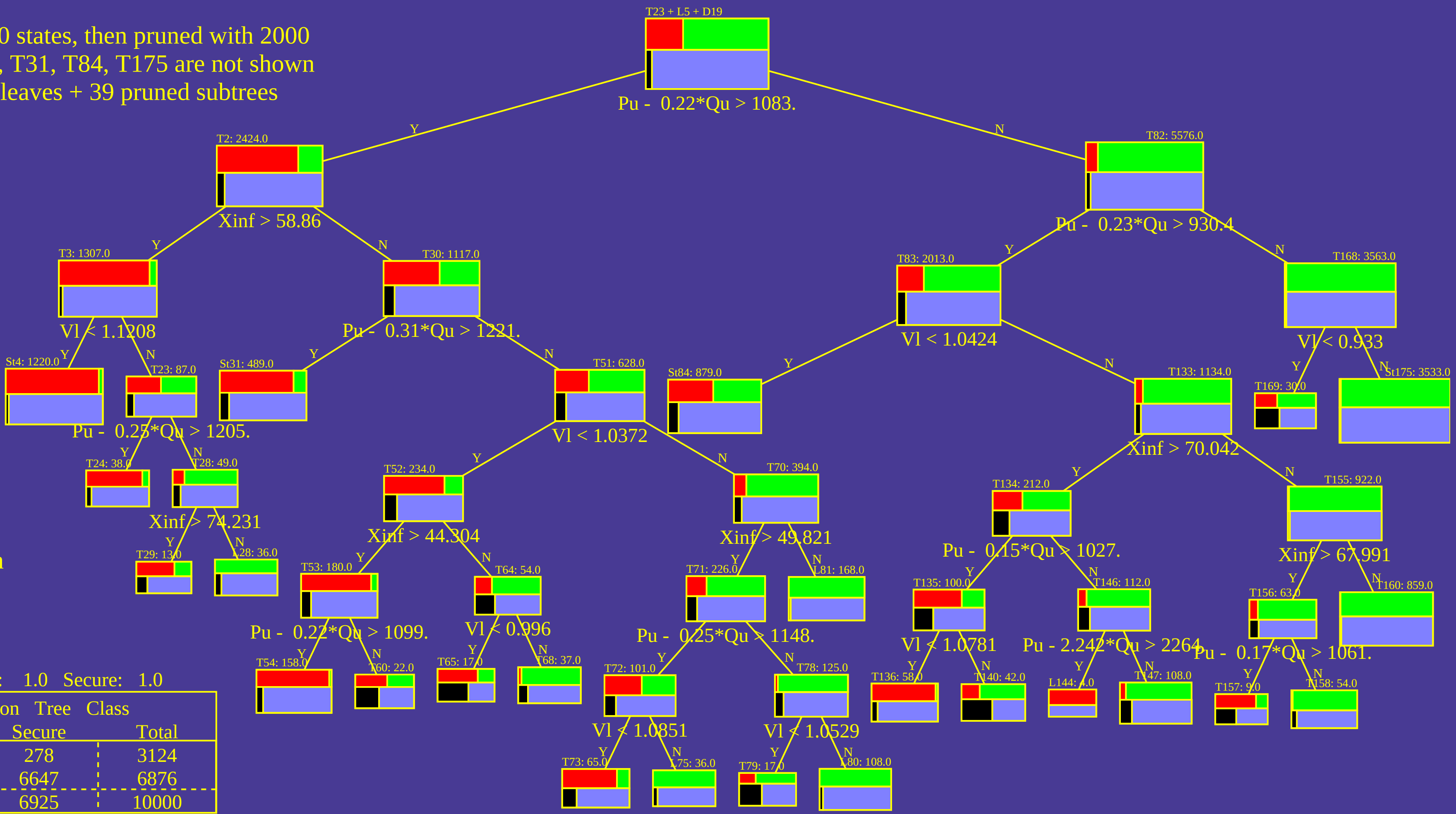
⇒ Investigation of possible ways to reduce threshold variance

Experimental setup

- Synthetic problem (Electric power systems transient stability)
- 6 attributes, 2 classes (stable/unstable), 20,000 random states
- Learning samples picked randomly among 10,000 first states
- Asymptotic values determined on the whole data base
- Experimental study of the discretization variance
  (only of the most informative attribute)
- Evaluation of bias and standard deviations of thresholds
  (other quantities, see paper)
- Repeated from small (N=50) to fairly large sample sizes (N=3000)

Example decision tree

Fully grown with 8000 states, then pruned with 2000
Subtrees at nodes : T4, T31, T84, T175 are not shown
61 test nodes + 23 leaves + 39 pruned subtrees



Learning set classification

🟥 Insecure: 2440

🟩 Secure: 5560

Test set classification.

Non detection costs : Insecure: 1.0 Secure: 1.0

Reference	Decision Tree Class		
Security	Insecure	Secure	Total
Insecure	2846	278	3124
Secure	229	6647	6876
Total	3075	6925	10000

Crisp discretization and its variance

Discretization (enumerative brute force) :

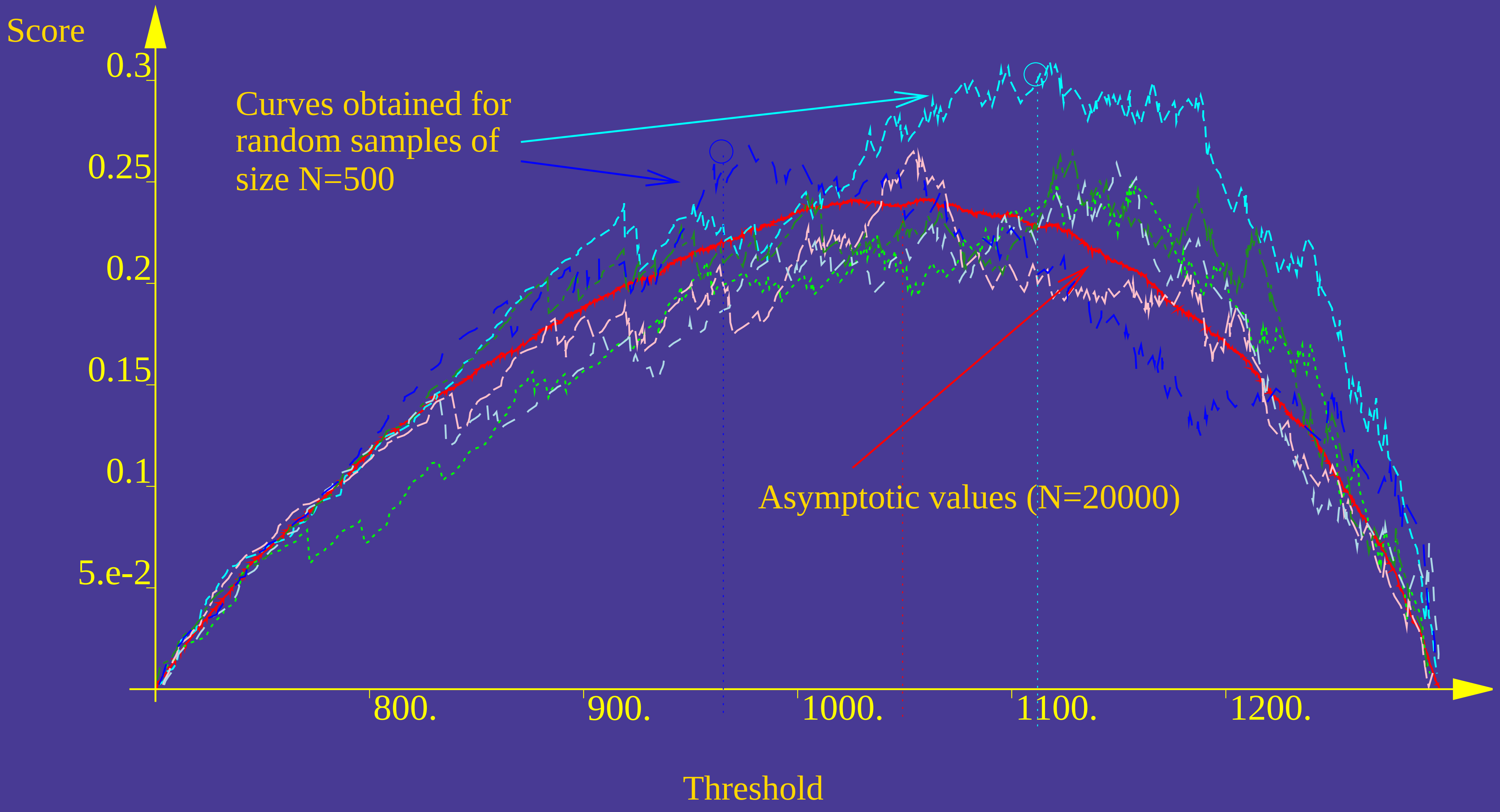
Given a classified sample S=$\{o_1, \ldots, o_N\}$

1. sort S by increasing order of values of $a(\,\cdot\,)$
2. for all pairs of values a_i, a_{i+1} compute score of test $T(o)\ : a(o) < \frac{a_i + a_{i+1}}{2}$?
3. return threshold corresponding to the maximum score

Information quantity provided by question $T(o)\ : a(o) < a_{th}$? on classes.

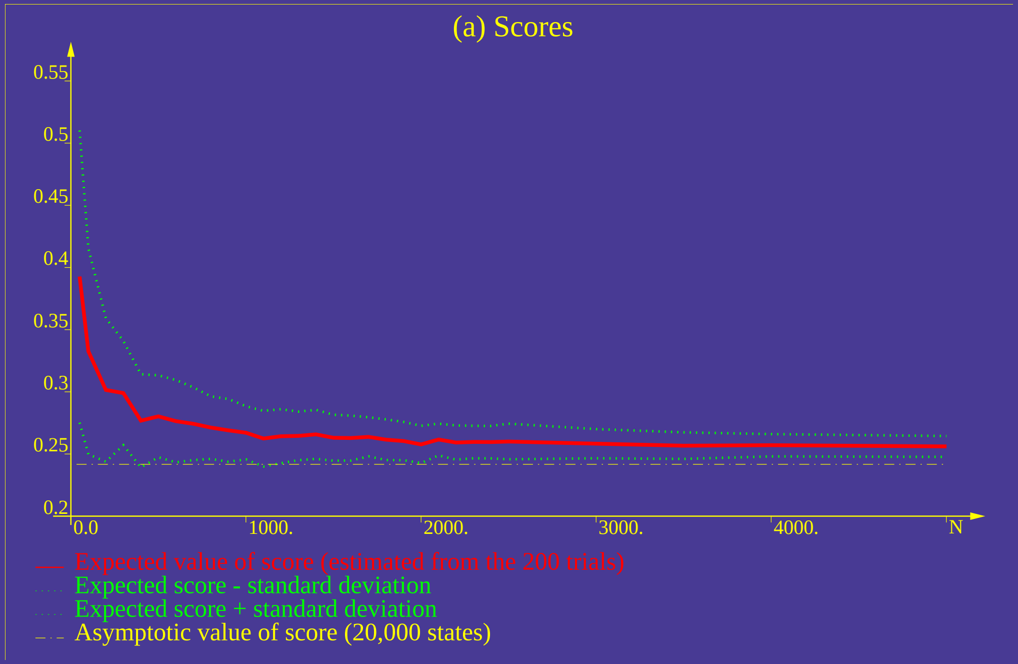
$$\widehat{H}_C = -\sum_j \frac{N_{.j}}{N} \log_2 \frac{N_{.j}}{N} \quad ; \quad \widehat{H}_T = -\sum_i \frac{N_{i.}}{N} \log_2 \frac{N_{i.}}{N} \quad ; \quad \widehat{I}_C^T = -\sum_i \sum_j \frac{N_{ij}}{N} \log_2 \frac{N_{i.}N_{.j}}{N_{ij}}.$$

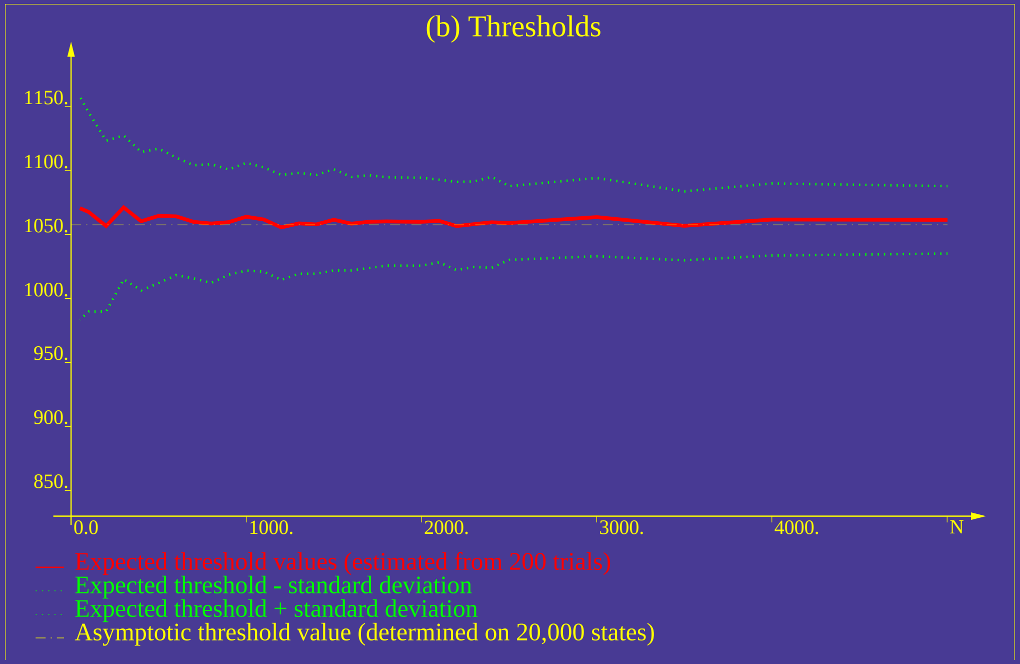
$\Rightarrow$ symmetric, normalized score measure : $C_C^T \triangleq \frac{2I_C^T}{H_C + H_T}$



Curves obtained for random samples of size N=500

Asymptotic values (N=20000)





Scores

- High bias for $N < 1000$
- Large variance for $N < 500$
- Small variance for $N > 2000$

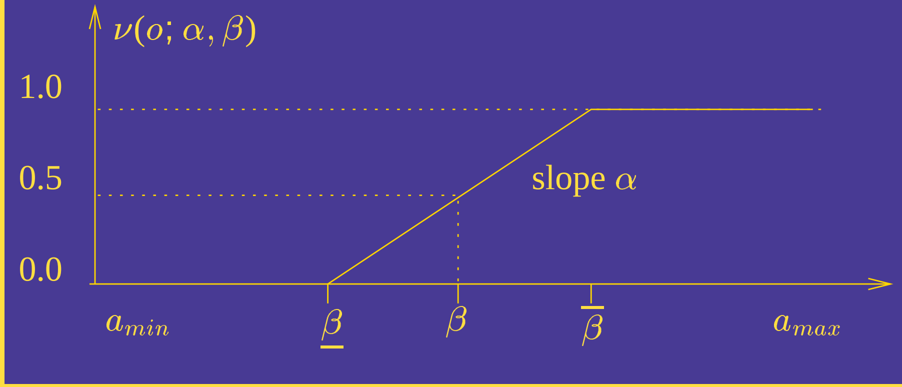
Thresholds

- No bias for all sample sizes
- Large variance for $N < 500$
- Variance remains large

Fuzzy discretization and its variance

Fuzzy discretization (two level enumerative brute force)

Search for combination of two thresholds $\underline{\beta}$, $\overline{\beta}$ maximizing $\mathcal{B}_\xi(\nu)$



Evaluate variance of β, $\overline{\beta}$ and $\underline{\beta}$

Compare with crisp discretization

$\Rightarrow$ Very small variance of β

For $N < 200$, large variance of $\overline{\beta}$ and $\underline{\beta}$

For $N > 500$, small variance of $\overline{\beta}$ and $\underline{\beta}$

Normalized fuzzy KS measure $\mathcal{B}_\xi(\nu) \triangleq \dfrac{\sqrt[\xi]{2}\,\left|\langle\nu\rangle_{S\cap C}-\langle\nu\rangle_{S\cap\overline{C}}\right|}{\sqrt[\xi]{\langle|2\,\nu-1|^\xi\rangle_{S\cap C}+\langle|2\,\nu-1|^\xi\rangle_{S\cap\overline{C}}}}$,

where $\langle\nu\rangle_A \triangleq \dfrac{\sum_{o\in A}\mu_A(o)\nu(o)}{\sum_{o\in A}\mu_A(o)}$.

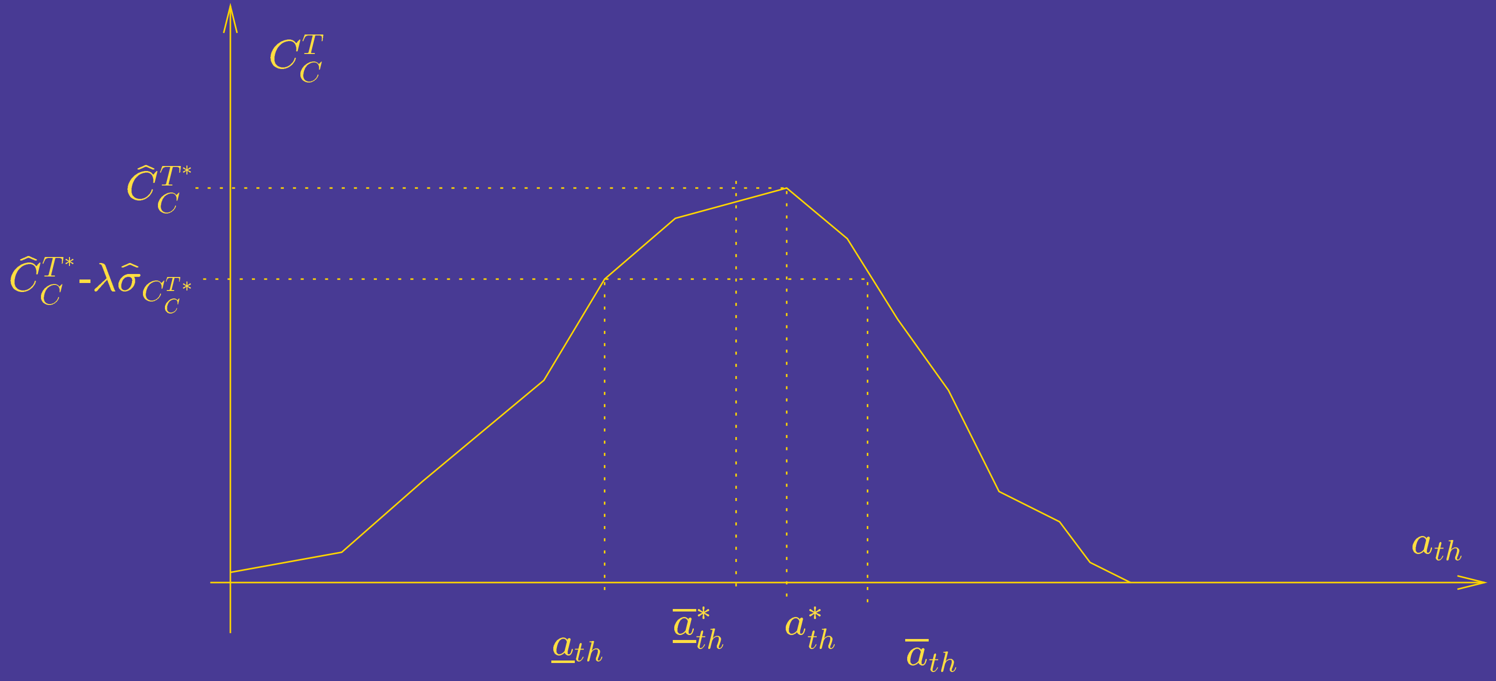
NB. Normalization is necessary to yield fuzzy discretization

NB. In our experiment class C is crisp

Variance reduction of crisp thresholds

Idea :

- try to estimate threshold uncertainty from learning sample : $a^*_{th} \in (\underline{a}_{th} \ldots \overline{a}_{th})$
- instead of a^*_{th} use $\underline{\overline{a}}^*_{th} = \frac{\underline{a}_{th} + \overline{a}_{th}}{2}$ as threshold

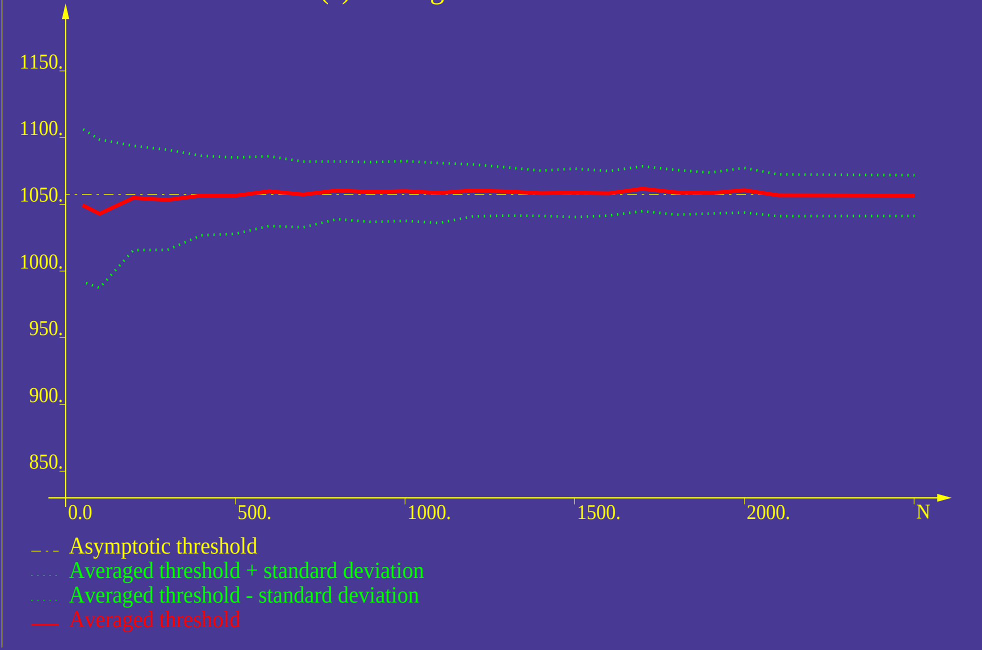


Estimator of score standard deviation :

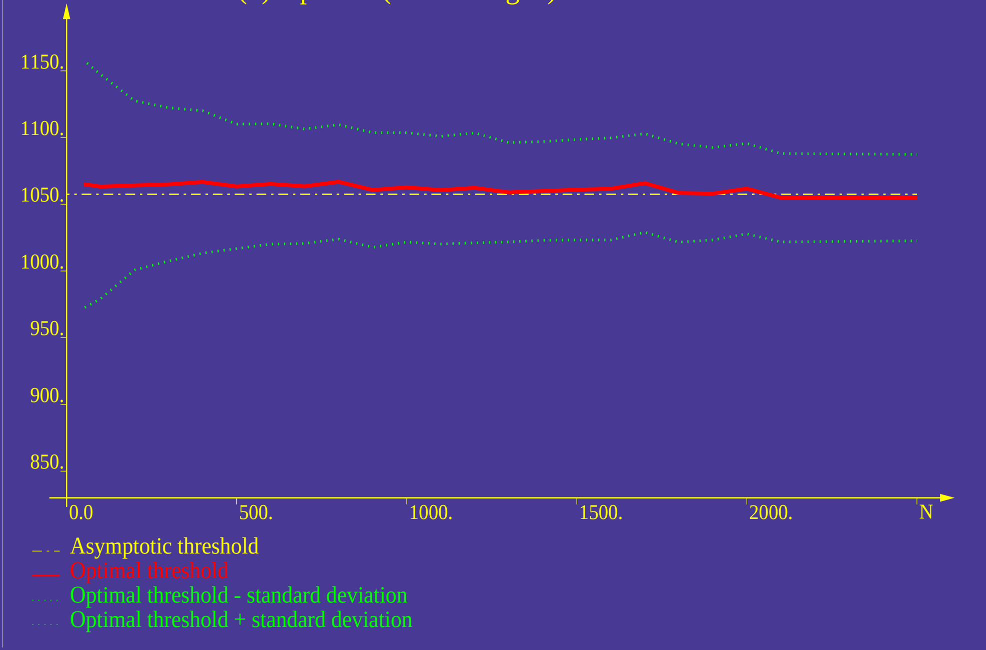
$$\widehat{\sigma}_{C^T_C} = \sqrt{\left(\frac{\widehat{C}^T_C}{N\widehat{I}^T_C}\right)^2 \sum_i \sum_j N_{ij}\left(\log_2 N_{ij} + \left(\frac{\widehat{C}^T_C}{2} - 1\right)\log_2 (N_{i.}N_{.j}) + (1 - \widehat{C}^T_C)\log_2 N^2\right)}$$

Comparison of the two scores

(a) Averaged threshold values



(b) Optimal (non averaged) threshold values



Conclusion :

50% reduction in threshold variance, with negligible computational overhead

⇒ less effective than fuzzy discretization, but very nice indeed !

Bayesian discretization

- Another approach to soft discretization
- Idea
  - Compute posterior probabilities of all thresholds, given S
  - Average all thresholds according to their posterior probability $\Rightarrow$ transition region
  - See paper for details

Principle

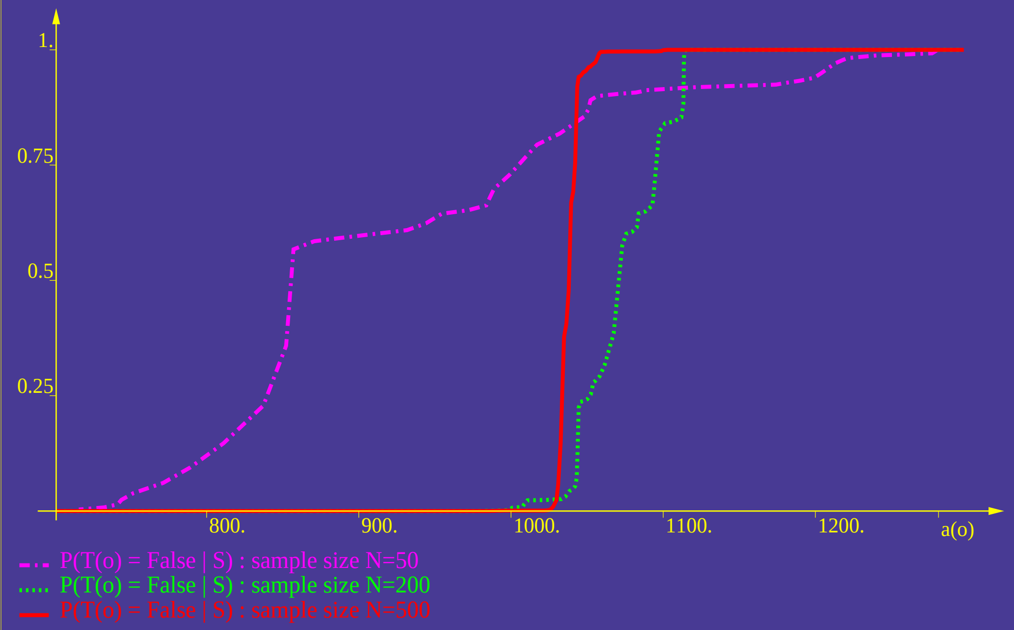
If a_{th} provides I_T^C in S of size N

$$\Rightarrow P(a_{th}|S) \propto \exp\left(N I_T^C\right)$$

Thus,

$$P(T(o){=}False|S){=} \int_{a_{min}}^{a(o)} P(a_{th}|S)da_{th}$$

$$P(T(o){=}False|S){=}\frac{\int_{a_{min}}^{a(o)} \exp\left(N I_T^C\right)da_{th}}{\int_{a_{min}}^{a_{max}} \exp\left(N I_T^C\right)da_{th}}$$



Comments

- Bayesian transition regions can be obtained as a biproduct of crisp discretization
- Needs further improvements (smoothing)
- Main differences with fuzzy discretization
  - Interpretation (of course ?)
  - Asymptotic behavior ($N \to \infty$)

    $\Rightarrow$ Fuzzy transition regions stabilize to class overlap region

    $\Rightarrow$ Bayesian transition regions stabilize to crisp threshold

NB. Fuzzy and Bayesian approaches may also be combined…

Summary

- Threshold variance of crisp discretization is often very high

⇒ Methods should be improved to reduce variance

- Crisp discretization can be improved at low cost
- Fuzzy and Bayesian approaches may be used to
  - Provide soft thresholds
  - Reduce variance
- Further work is neeeded to
  - Evaluate effect on machine learning performance
  - Smoothen Bayesian thresholds
  - Improve fuzzy discretization (computational, variance)
  - Consider combinations of fuzzy and Bayesian approaches