



HAL
open science

Embedded Variable Selection in Classification Trees

Servane Gey, Tristan Mary-Huard

► **To cite this version:**

Servane Gey, Tristan Mary-Huard. Embedded Variable Selection in Classification Trees. GfKI 2011 : Joint Conference of the German Classification Society (GfKI) and the German Association for Pattern Recognition (DAG), Aug 2011, Francfort, Germany. n.p. hal-01019767

HAL Id: hal-01019767

<https://hal.science/hal-01019767v1>

Submitted on 6 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Embedded Variable Selection in Classification Trees

Servane Gey¹, Tristan Mary-Huard²

¹MAP5, UMR 8145, Université Paris Descartes, Paris, France

²UMR AgroParisTech/INRA 518, Paris, France



Introduction

- ★ Binary classification setting
- ★ Model and variable selection in classification
- ★ Classification tree

Variable selection for CART

- ★ Classes of classification trees
- ★ Theoretical results
- ★ Comparison with practice.

Binary classification

Binary classification

Prediction of the unknown label Y (0 or 1) of an observation X .

⇒ Use training sample $D = (X_1, Y_1), \dots, (X_n, Y_n) \stackrel{i.i.d}{\sim} \mathbb{P}$ to build a classifier \hat{f}

$$\begin{aligned}\hat{f}: \mathcal{X} &\rightarrow \{0, 1\} \\ X &\mapsto \hat{Y}.\end{aligned}$$

Quality assessment

★ Classification risk and loss : Quality of the resulting classifier \hat{f}

$$\begin{aligned}L(\hat{f}) &= \mathbb{P}(\hat{f}(X) \neq Y|D) \\ \ell(\hat{f}, f^*) &= L(\hat{f}) - L(f^*)\end{aligned}$$

★ Average loss : Quality of the classification algorithm

$$\mathbb{E}_D[\ell(\hat{f}, f^*)]$$

Remark : All these quantities depend on \mathbb{P} that is unknown.

Basics of Vapnik Theory : structural risk minimization (SRM)

Consider and collection of classes of classifiers $\mathcal{C}_1, \dots, \mathcal{C}_M$. Define

$$\bar{f}_m = \arg \min_{f \in \mathcal{C}_m} L(f), \quad \hat{f}_m = \arg \min_{f \in \mathcal{C}_m} L_n(f), \quad \hat{f} = \arg \min_m \left[L_n(\hat{f}_m) + \alpha \frac{V_{\mathcal{C}_m}}{n} \right]$$

★ Class complexity

If $\mathcal{C}_1, \dots, \mathcal{C}_M$ have finite VC dimensions $V_{\mathcal{C}_1}, \dots, V_{\mathcal{C}_M}$, then

$$\mathbb{E}_D [\ell(\hat{f}, f^*)] \leq C \left\{ \inf_m \left(\ell(\bar{f}_m, f^*) + K \sqrt{\frac{V_{\mathcal{C}_m}}{n}} \right) \right\} + \frac{\lambda}{n} \quad (\text{Vapnik, 1998}).$$

★ Classification task complexity (Margin Assumption)

If there exists $h \in]0; 0.5[$ such that

$$\mathbb{P} (|\eta(x) - 1/2| \leq h) = 0, \quad \text{with } \eta(x) = \mathbb{P}(Y = 1 | X = x)$$

then

$$\mathbb{E}_D [\ell(\hat{f}, f^*)] \leq C \left\{ \inf_m \left(\ell(\bar{f}_m, f^*) + K' \left(\frac{V_{\mathcal{C}_m}}{n} \right) \right) \right\} + \frac{\lambda'}{n} \quad (\text{Massart \& Nédélec, 2006}).$$

Assume that $X \in \mathbb{R}^p$. Define

$$\bar{f}_{m(k)} = \arg \min_{f \in \mathcal{C}_{m(k)}} L(f), \quad \hat{f}_{m(k)} = \arg \min_{f \in \mathcal{C}_{m(k)}} L_n(f)$$

★ Variable selection

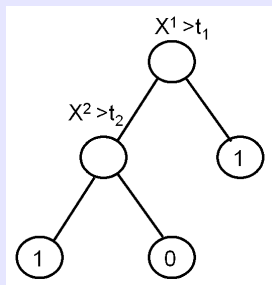
Choose \hat{f} such that

$$\hat{f} = \arg \min_{m(k)} \left[L_n(\hat{f}_{m(k)}) + \alpha \frac{V_{\mathcal{C}_{m(k)}}}{n} + \alpha' \log \binom{p}{k} \right]$$

Then (under strong margin assumption)

$$\mathbb{E}_D [\ell(\hat{f}, f^*)] \leq C \log(p) \left\{ \inf_{m(k)} \left(\ell(\bar{f}_{m(k)}, f^*) + K' \left(\frac{V_{\mathcal{C}_{m(k)}}}{n} \right) \right) \right\} + \frac{\lambda}{n}$$

(Massart, 2000, Mary-Huard et al., 2007)



General strategy

Choose

$$\hat{f} = \arg \min_T L_n(f_T) + \alpha \frac{|T|}{n}$$

Heuristic approach (CART, Breiman, 1984)

★ Find a tree T_{\max} such that $L_n(f_{T_{\max}}) = 0$,

★ Prune T_{\max} using criterion :

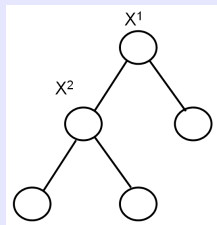
$$\hat{f} = \arg \min_{T \subseteq T_{\max}} L_n(f_T) + \alpha \frac{|T|}{n}$$

Definitions

Consider a tree T_{cl} with

- a given configuration c ,
- a given list ℓ of associated variables.

Remark : A same variable may be associated with several nodes.



Class of tree classifiers

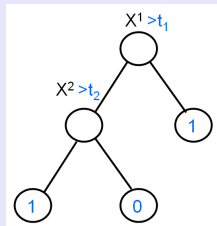
Define

$$\mathcal{C}_{cl} = \{f/f \text{ based on } T_{cl}\},$$

$$H_{cl} = \text{VC log-entropy of class } \mathcal{C}_{cl},$$

$$\bar{f}_{cl} = \underset{f \in \mathcal{C}_{cl}}{\operatorname{argmin}} L(f),$$

$$\hat{f}_{cl} = \underset{f \in \mathcal{C}_{cl}}{\operatorname{argmin}} L_n(f).$$



Remark : Two classifiers $f, f' \in \mathcal{C}_{cl}$ only differ in their thresholds and labels.

Proposition

Assume that strong margin assumption is satisfied. For all $C > 1$, there exist positive constants K^1 and K^2 depending on C such that

$$\mathbb{E}_D[\ell(\widehat{f}_{cl}, f^*)] \leq C \left\{ \ell(\bar{f}_{cl}, f^*) + K^1 \left(\frac{|T_{cl}| \log(2n)}{n} \right) \right\} + \frac{K^2}{n} .$$

Idea of proof

- ★ Show that $E[H_{cl}] \leq |T_{cl}| \log(2n)$,
- ★ Apply general theory (Koltchinskii, 2006).

To take into account variable selection in the penalized criterion, one needs to count the number of classes sharing the same a priori complexity.

★ **Parametric case** (Logistic regression, LDA,...)

- One parameter per variable,
- 2 classes with classifiers based on k variables have the same a priori complexity,
⇒ $\binom{p}{k}$ classes of a priori complexity k .

★ **Classification trees**

- One parameter per internal node (threshold),
- 2 classes \mathcal{C}_{cl} and $\mathcal{C}_{c'l'}$ such that $|T_{cl}| = |T_{c'l'}|$ have the same a priori complexity
⇒ **Count the number of classes based on trees of size k !**

Combinatorics for variable selection

A tree T_{cl} is defined by

- a configuration,
- a list of variables associated with each node.

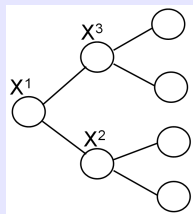
★ Number of configurations of size k :

$$N_c^k = \frac{1}{k} \binom{2k-2}{k-1}$$

★ Number of variable lists of size k :

- the list is ordered : $\{1, 2, 3\} \neq \{2, 1, 3\}$,
- variables are selected with replacement : $\{1, 2, 1\}$.

$$\Rightarrow N_\ell^k = p^{k-1} \text{ instead of } \binom{p}{k} !$$



★ Number of classes based on trees of size $|T_{cl}| = k$:

$$N^k = N_c^k \times N_\ell^k = \frac{1}{k} \binom{2k-2}{k-1} \times p^{k-1}$$

$$\Rightarrow \log(N^k) \leq \lambda |T_{cl}| \log(p)$$

Combinatorics for variable selection

A tree T_{cl} is defined by

- a configuration,
- a list of variables associated with each node.

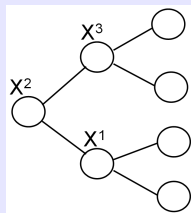
★ Number of configurations of size k :

$$N_c^k = \frac{1}{k} \binom{2k-2}{k-1}$$

★ Number of variable lists of size k :

- the list is ordered : $\{1, 2, 3\} \neq \{2, 1, 3\}$,
- variables are selected with replacement : $\{1, 2, 1\}$.

$$\Rightarrow N_\ell^k = p^{k-1} \text{ instead of } \binom{p}{k} !$$



★ Number of classes based on trees of size $|T_{cl}| = k$:

$$N^k = N_c^k \times N_\ell^k = \frac{1}{k} \binom{2k-2}{k-1} \times p^{k-1}$$

$$\Rightarrow \log(N^k) \leq \lambda |T_{cl}| \log(p)$$

Combinatorics for variable selection

A tree T_{cl} is defined by

- a configuration,
- a list of variables associated with each node.

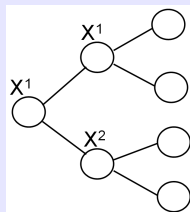
★ Number of configurations of size k :

$$N_c^k = \frac{1}{k} \binom{2k-2}{k-1}$$

★ Number of variable lists of size k :

- the list is ordered : $\{1, 2, 3\} \neq \{2, 1, 3\}$,
- variables are selected with replacement : $\{1, 2, 1\}$.

$$\Rightarrow N_\ell^k = p^{k-1} \text{ instead of } \binom{p}{k} !$$



★ Number of classes based on trees of size $|T_{cl}| = k$:

$$N^k = N_c^k \times N_\ell^k = \frac{1}{k} \binom{2k-2}{k-1} \times p^{k-1}$$

$$\Rightarrow \log(N^k) \leq \lambda |T_{cl}| \log(p)$$

Proposition

Assume that strong margin assumption is satisfied. If

$$\hat{f} = \arg \min_{c, \ell} (L_n(\hat{f}_{c\ell}) + \text{pen}(c, \ell)),$$
$$\text{where } \text{pen}(c, \ell) = C_h^1 \frac{|T_{c\ell}| \log(2n)}{n} + C_h^2 \frac{|T_{c\ell}| \log(\rho)}{n}$$

with constants C_h^1, C_h^2 depending on h appearing in the margin condition, then there exist positive constants C, C', C'' such that

$$\mathbb{E}_D[l(\hat{f}, f^*)] \leq C \log(\rho) \left\{ \inf_{c, \ell} \left\{ \ell(\bar{f}_{c\ell}, f^*) + C' \left(\frac{|T_{c\ell}| \log(2n)}{n} \right) \right\} \right\} + \frac{C''}{n}.$$

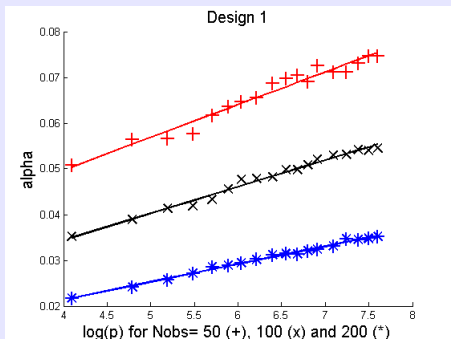
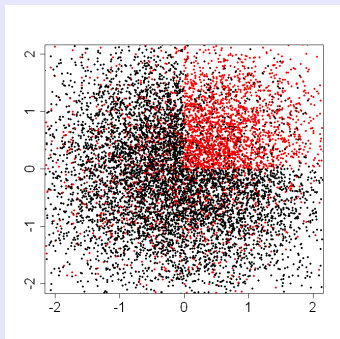
Remark :

$$\text{Theory : } \text{pen}(c, \ell) = (a_n + b_n \log(\rho)) |T_{c\ell}| = \alpha_{p,n} |T_{c\ell}|$$

$$\text{Practice (CART) : } \text{pen}(c, \ell) = \alpha_{CV} |T_{c\ell}|$$

Does α_{CV} match $\alpha_{p,n}$?

Illustration on simulated data (1)

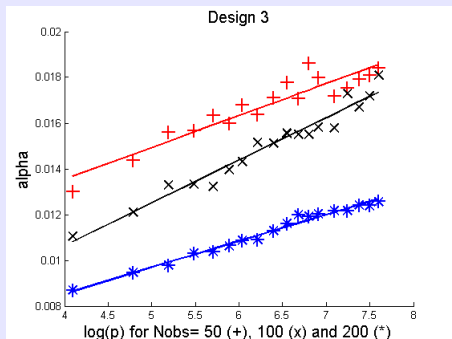
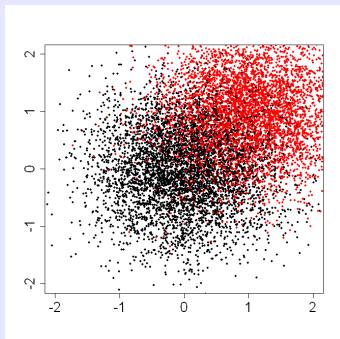


- Variables X^1, \dots, X^p are independent,
- If $X^1 > 0$ and $X^2 > 0$ $P(Y = 1) = q$, otherwise $P(Y = 1) = 1 - q$

Remark : Easy case

- The Bayes classifier belongs to the collection of classes,
- Strong margin assumption is satisfied.

Illustration on simulated data (2)



- $P(Y = 1) = 0.5$
- For $j = 1, 2$, $X^j | Y = 0 \hookrightarrow \mathcal{N}(0, \sigma^2)$ and $X^j | Y = 1 \hookrightarrow \mathcal{N}(1, \sigma^2)$,
- Additional variables are independent and non-informative.

Remark : Difficult case

- The Bayes classifier does NOT belong to the collection of classes,
- Strong margin assumption is NOT satisfied.

Model selection for tree classifiers :

- Already investigated (Nobel 02, Gey & Nedelec 06, Gey 10),
- Variable selection not investigated so far.
- Pruning step now validated from this point of view.

Theory vs practice

- Theory : exhaustive search,
- Practice : forward strategy,
- Nonetheless theoretical results are informative !

Extension

- In this talk : strong margin assumption
- Can be extended to less restrictive margin assumption
- Manuscript on arXiv.org :

<http://arxiv.org/abs/1108.0757>

- Breiman L., Friedman J., Olshen R. & Stone, C.** (1984) *Classification And Regression Trees*, Chapman & Hall.
- Gey S. & Nédélec E.** (2005) *Model selection for CART regression trees*, IEEE Trans. Inform. Theory, 51, 658–670.
- Koltchinskii, V.** (2006) *Local Rademacher Complexities and Oracle Inequalities in Risk Minimization*, Annals of Statistics, 34, 2593–2656.
- Mary-Huard T., Robin S. & Daudin J.-J.** (2007) *A penalized criterion for variable selection in classification*, J. of Mult. Anal., 98, 695–705.
- Massart P.** (2000) *Some applications of concentration inequalities to statistics*, Annales de la Faculté des Sciences de Toulouse.
- Massart P. & Nédélec E.** (2006) *Risk Bounds for Statistical Learning*, Annals of Statistics, 34, 2326–2366.
- Nobel A.B.** (2002) *Analysis of a complexity-based pruning scheme for classification trees*, IEEE Trans. Inform. Theory, 48, 2362–2368.