On the Consistency Rate of Decision Tree Learning Algorithms

Qin-Cheng Zheng, Shen-Huan Lyu, Shao-Qun Zhang, Yuan Jiang, Zhi-Hua Zhou

{ zhengqc, lvsh, zhangsq, jiangy, zhouzh }@lamda.nju.edu.cn

Paper in a Nutshell

1. We found that the worst-case zero purity gain leads to a serious obstacle for consistency analysis of CART.
2. We found that using Influence as the impurity measure can always get a positive purity gain, but an Influence oracle to is required for generating a tree.
3. We propose the GridCART, an Influence-based CART, which not only can run practically but also is consistent with order $O(n^{-1/(d+2)})$.

Background

Heuristical algorithms maximize the *purity gain* to spit each node to two children.

ratio
impurity

$$(i^\star, s^\star) \in \arg\max_{i,s} \mathcal{G}(t) - w_{t_L}\mathcal{G}(t_L) - w_{t_R}\mathcal{G}(t_R)$$

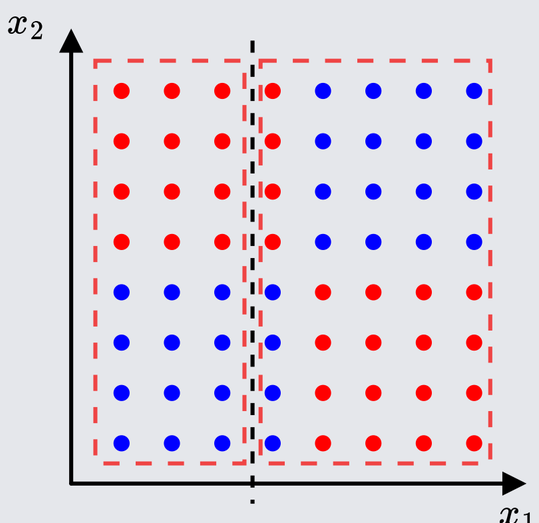
children

This type of algorithm succeeds in many real-world tasks, but the consistency is *far from clear*.

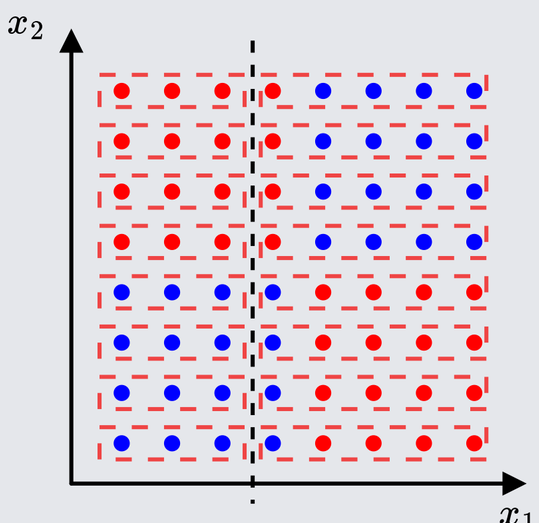
CART vs InfCART

Algorithm	Impurity Measure
CART	$\mathcal{G}(t) \triangleq G\left(\mathbb{E}_{\mathbf{x}(j\neq i)}\left[\mathbb{E}_{\mathbf{x}^{(i)}}[Y \mid \mathbf{x} \in t]\right]\right)$
InfCART	$\mathcal{G}(t) \triangleq \mathbb{E}_{\mathbf{x}(j\neq i)}\left[G\left(\mathbb{E}_{\mathbf{x}^{(i)}}[Y \mid \mathbf{x} \in t]\right)\right]$

CART always gets zero purity gain in the worst case see (a), while InfCART (a variant) calculates the average purity gain on every line and gets a positive average purity gain see (b).



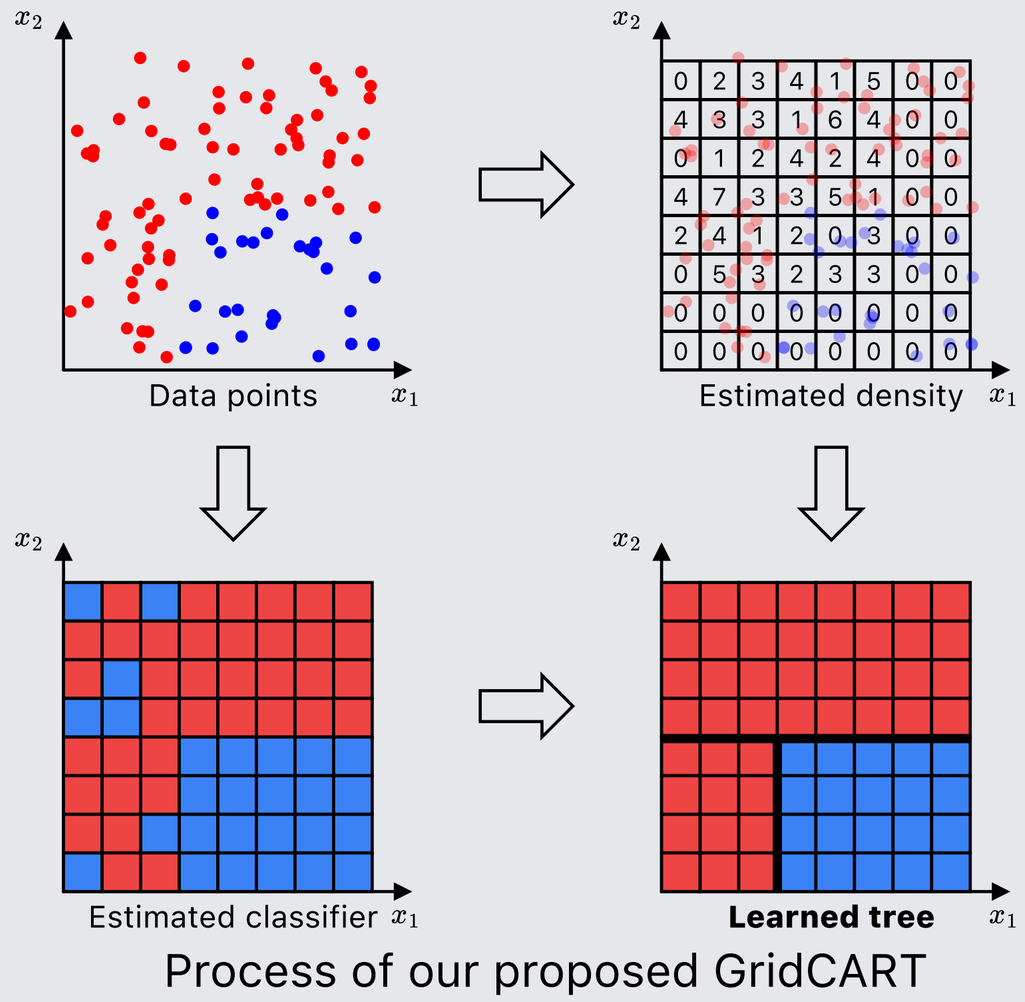
(a) The worst-case purity gain is zero for CART.



(b) The purity gain is always positive for InfCART.

Our Proposed GridCART

GridCART can not only be practical but also always obtains a positive purity gain. The key is to estimate the probability density, and further the Influence.



Process of our proposed GridCART

Theoretical Guarantees

Consistency rate for GridCART

Theorem 8 (informal). *Under certain assumptions, the expected excess error of tree T_{K_n} learned by GridCART has the following upper bound*

$$\mathbb{E}_{D_n}\left[R(T_{K_n})\right] - R^\star \leq \mathcal{O}\left(\frac{1}{K_n h_n^3} + h_n + \sqrt{\frac{1}{n h_n^d}}\right).$$

depth · cube size · sample size

Choosing $h_n = \Theta\left(n^{-1/(d+2)}\right)$, $K_n = \Omega(n^{4/(d+2)})$, we obtain a consistency rate of order $\mathcal{O}\left(n^{-1/(d+2)}\right)$.

GridCART can reach a consistency rate of order $O(n^{-1/(d+2)})$ under certain assumptions. However, the consistency rate of CART is far from clear.

Algorithm	Advantage	Disadvantage
CART	Empirically practical	*Worst-case zero purity gain*
InfCART	Positive average purity gain	*An Influence oracle required*