決定木アンサンブルの分岐条件の共有化

情報認識学研究室
櫻田 健斗

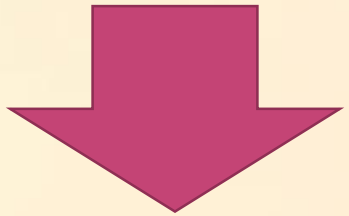
背景

- 機械学習の学習アルゴリズムや予測アルゴリズムの需要の増加
- IoTの普及によるエッジコンピューティングへの期待

高速・省メモリの機械学習・予測アルゴリズム実装の必要性が高まる
→専用ハードウェア化(例：深層学習専用ハードウェア)

目的

- 勾配ブースティング木，ランダムフォレストは高性能な予測器
- そのままハードウェア化すると計算リソースを多く要求する



決定木アンサンブルをハードウェア上の使用計算リソースが
少なくなるように簡略化する

決定木アンサンブルの簡略化の既存研究

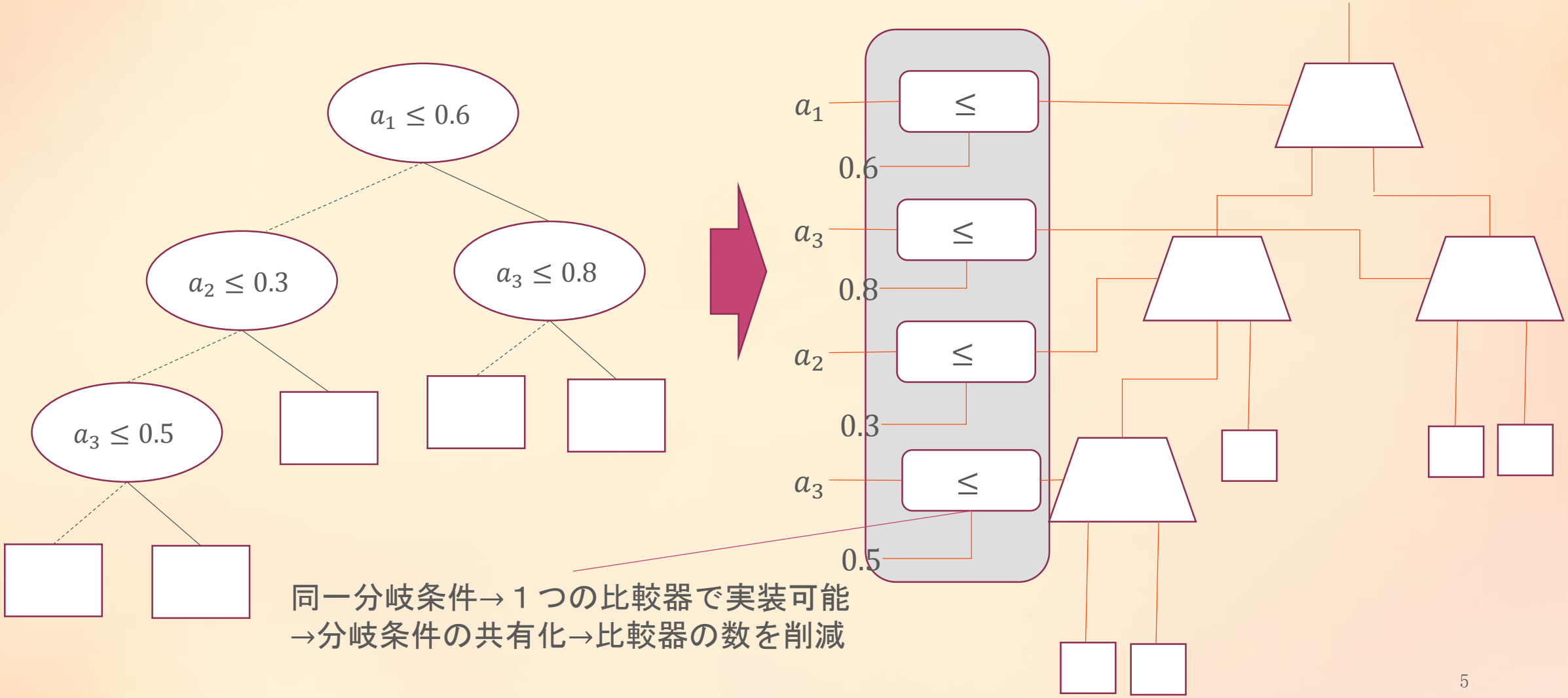
汎化性能の向上

- 決定木の枝刈り[Bradford *et al.*, *1998*]
- 決定木アンサンブルの枝刈り[Kulkarni, 2012]

高速・省スペースな実装

- 数値量子化[Markus *et al.*, 2018]
- 分岐条件の共有[Jinguji *et al.*, 2018]

分岐条件の共有化による計算リソース削減



同一分岐条件→１つの比較器で実装可能
→分岐条件の共有化→比較器の数を削減

提案手法

与えられた決定木アンサンブル$F[T_1, \ldots, T_K]: X \to Y$に対し，

属性ベクトル集合$D \subset X$を与え，

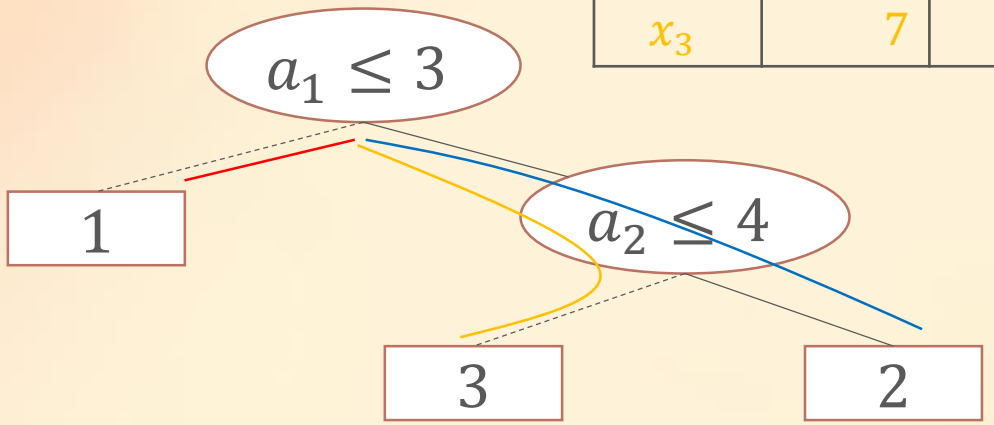
各属性ベクトル$\boldsymbol{x} \in D$の各決定木$T_K\ (i = 1, \ldots, K)$における

決定パスを変えないという条件の下に分岐条件を最大限共有する

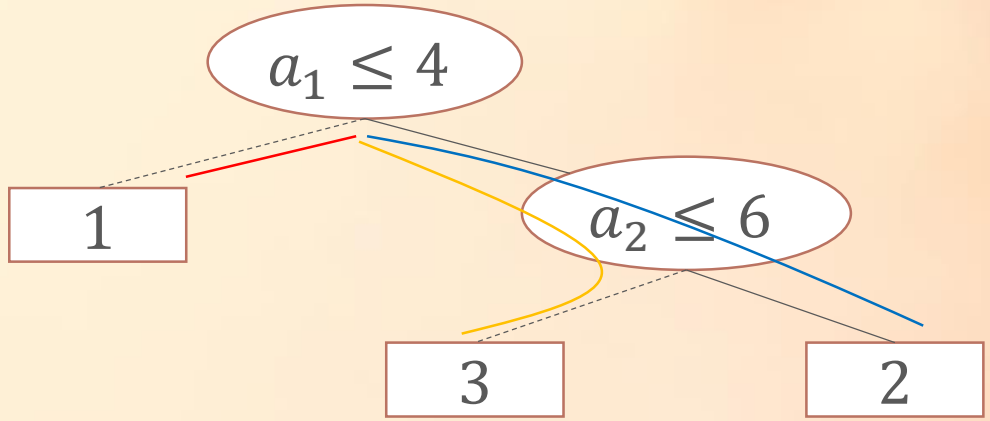
決定パス：データの根ノードから葉ノードまでに至る分岐ノードの列

属性	a_1	a_2
x_1	1	5
x_2	5	7
x_3	7	1

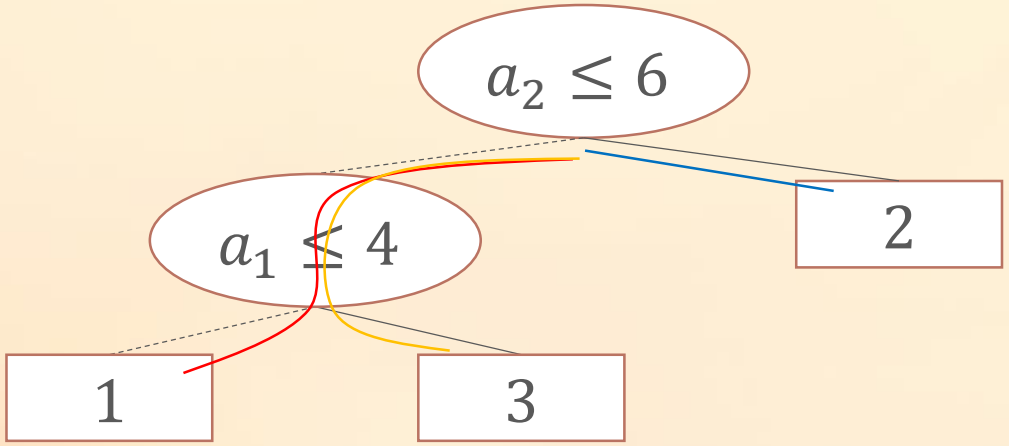
T_1



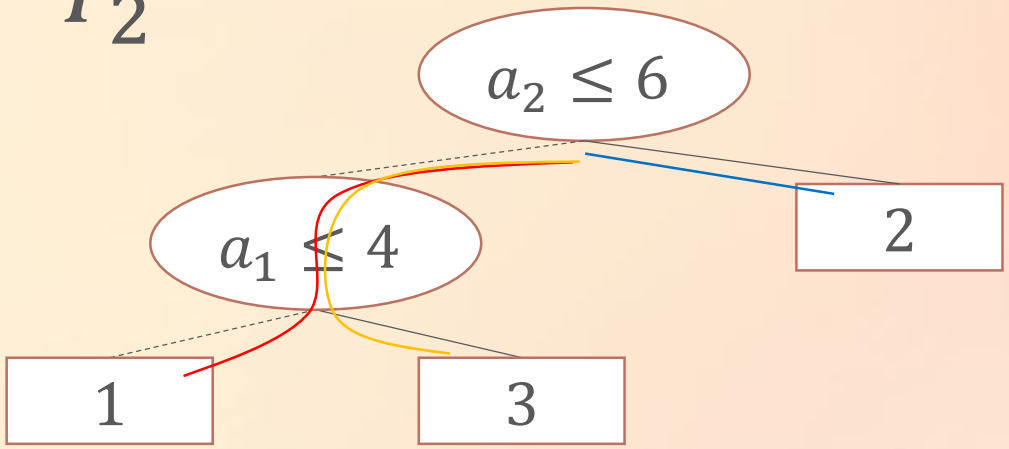
T_1'



T_2



T_2

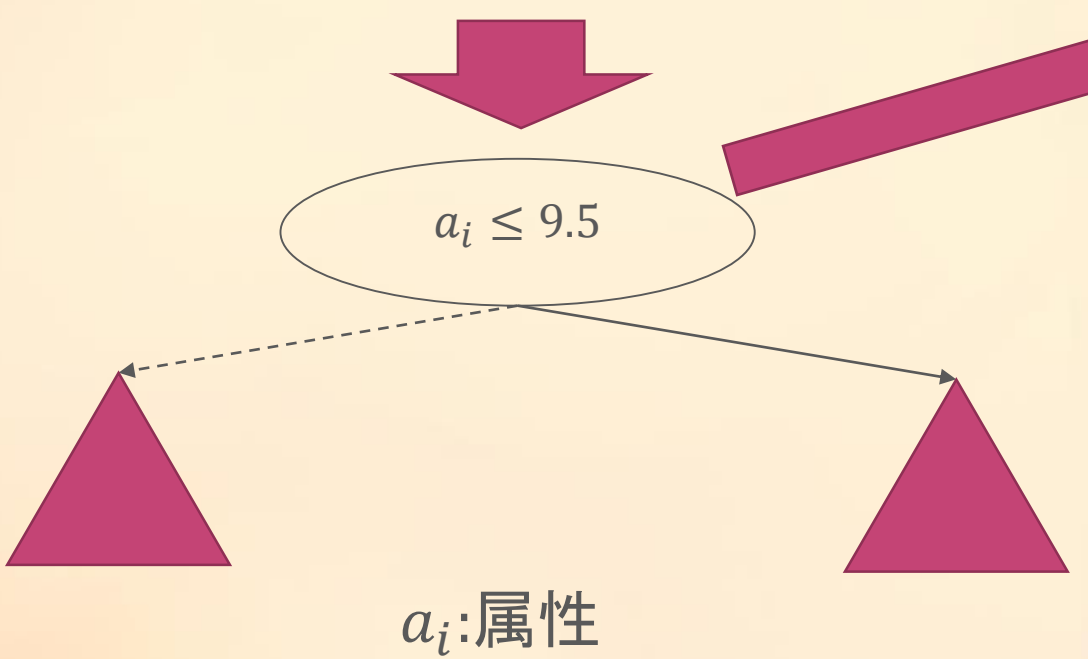


同じ決定パスを返すことができる閾値の区間

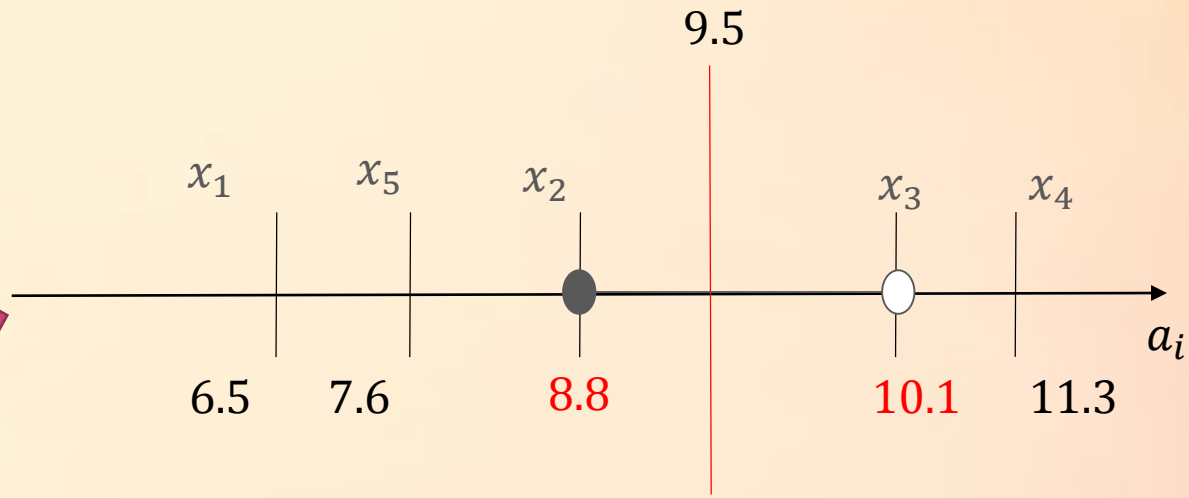
例

訓練データ
x_1, x_2, x_3, x_4, x_5

$a_i \leq 9.5$



a_i:属性



訓練データに対して同じ決定パスを返す区間
$[8.8, 10.1)$

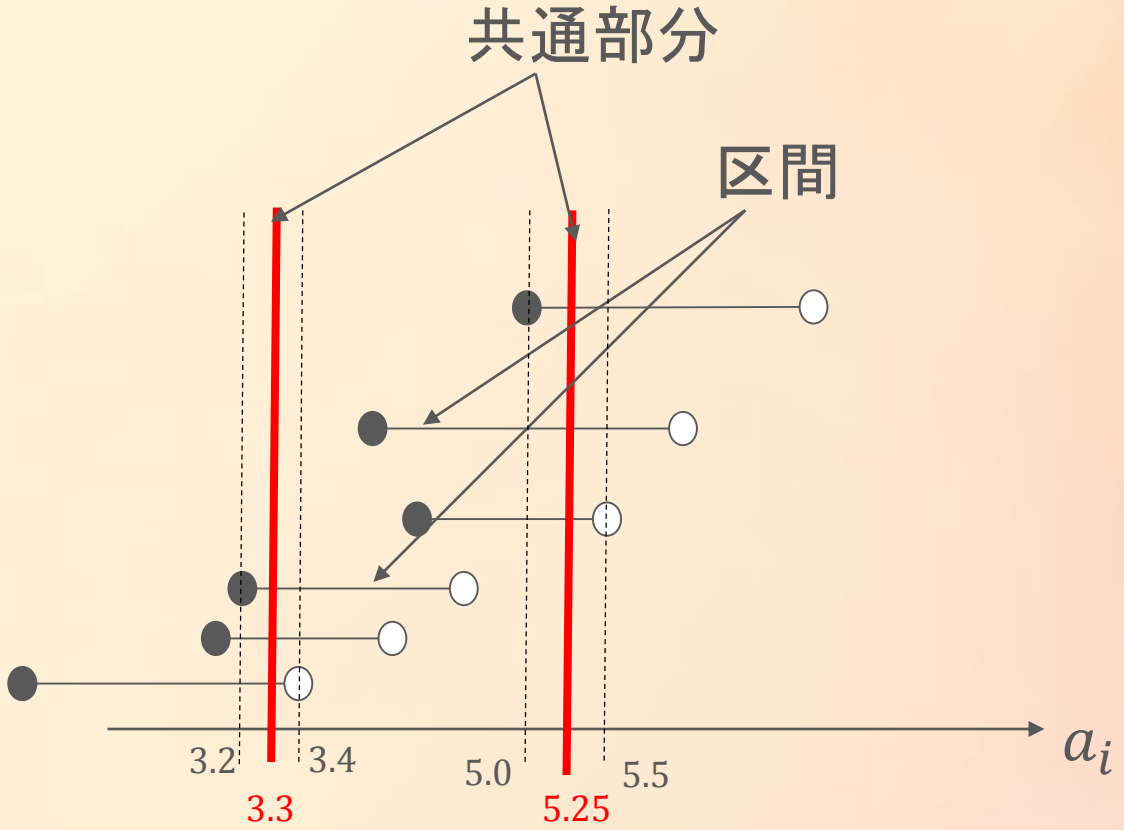
属性a_iについて全ての分岐条件から
求めて集める

全ての区間と共通部分を持つ最小集合 (全区間交差最小集合問題)

1. ある属性についての区間集合を上限についてソートする

2. ソートされた順で最初のk区間の共通部分を求める．共通部分が空集合になるまでkを大きくする．

3. 空集合になったら最初のk−1区間の共通部分である区間の中点を閾値の集合に入れる。



時間計算量：$O(N(n\log(n+1)+\log N)+d)$
N: 決定木の平均ノード数
n: 訓練データのサンプル数
d: 属性の次元

実験内容

1. Random Forestを用いた分類・回帰問題において効果の検証
2. Random Forestを分類問題へ適用した場合の，既存手法である k-meansを用いたクラスタリングによる閾値共有化法との効果の比較
3. 決定木アンサンブル(ExtraTrees, Adaboost, Gradient Boosting)を用いた分類問題において効果の検証

実験設定

- 決定木アンサンブルの実装:scikit-learn
- 木の本数：100
- データセット:UCI Machine Learning Repositoryより取得
- 訓練データ：テストデータ=4:1
  (ランダム分割．テストデータが別途提供されている場合はそれを使用)
- 実験結果：10回実行の平均

実験設定

評価項目

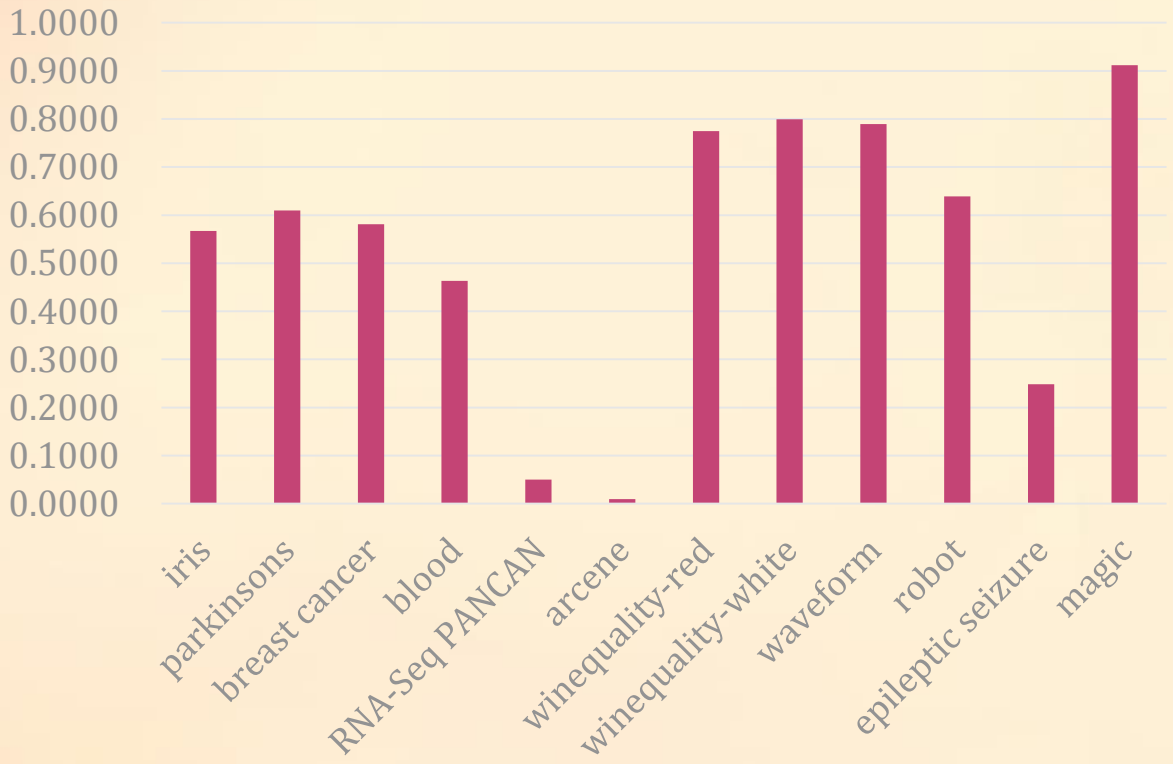
- 削減率: $1 - \dfrac{\text{異なる分岐条件数(簡略化後)}}{\text{異なる分岐条件数(簡略化前)}}$
- 予測精度比：$\dfrac{\text{テストデータの予測精度 (簡略化後)}}{\text{テストデータの予測精度 (簡略化前)}}$ （分類問題）
- RMSE比： $\dfrac{RMSE(\text{簡略化前})}{RMSE(\text{簡略化後})}$ （回帰問題）

分類問題データセット

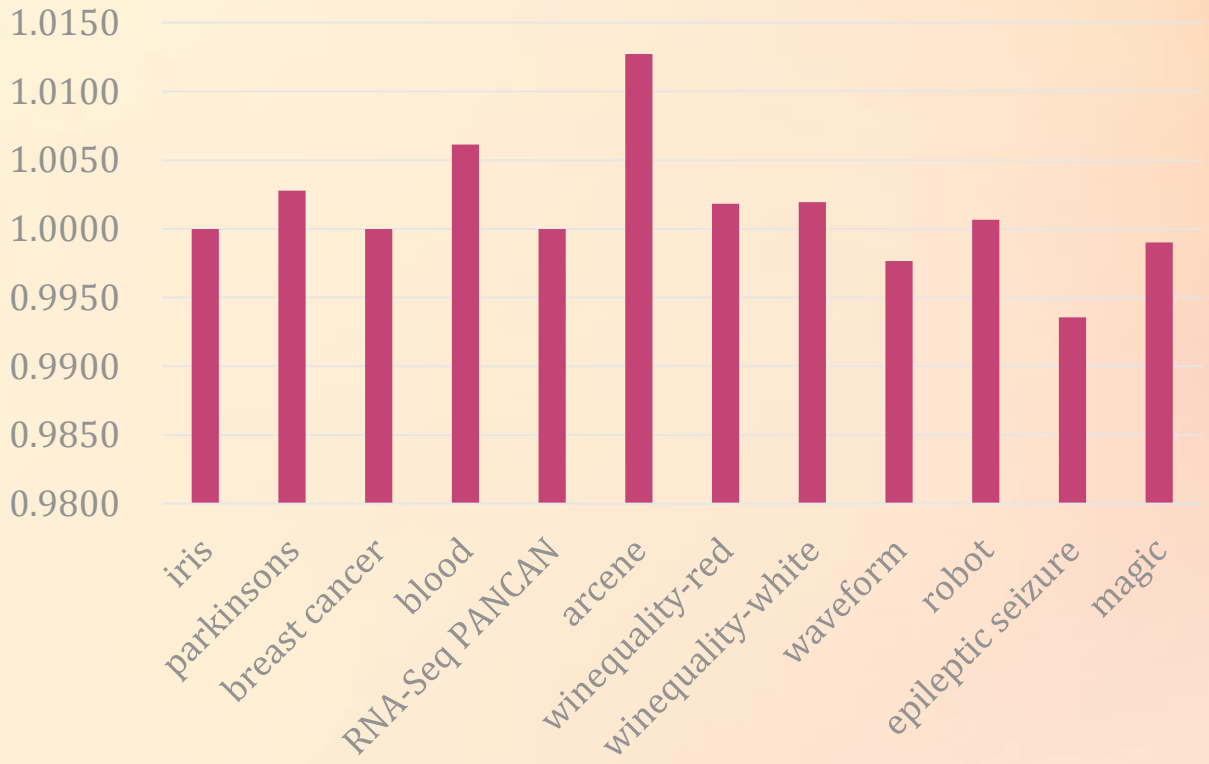
データセット名	サンプル数	属性の次元	クラス数
Iris	150	4	3
Parkinsons	195	22	2
Breast cancer	569	30	2
Blood	748	4	2
RNA-Seq PANCAN	801	20531	5
Arcene	900	10000	5
Winequality-red	1599	11	11
Winequality-white	4898	11	11
Waveform	5000	40	3
Robot	5456	24	4
Eplileptic seizure	11500	178	5
magic	19020	10	2

Random Forest 分類問題 実験結果

削減率



予測精度比



分類問題 実験結果

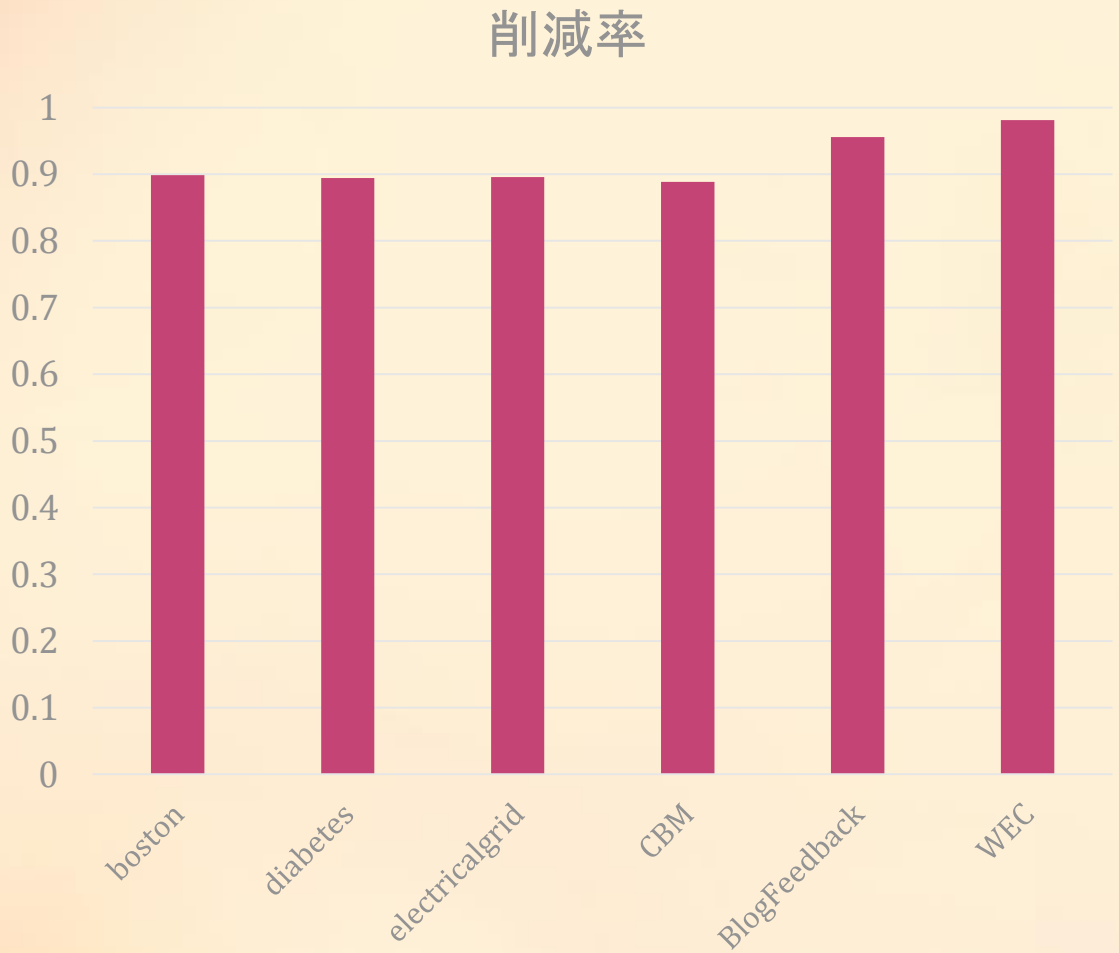
削減率が高い
magic, winequalityなど
→属性の次元が低い，サンプル数が多い

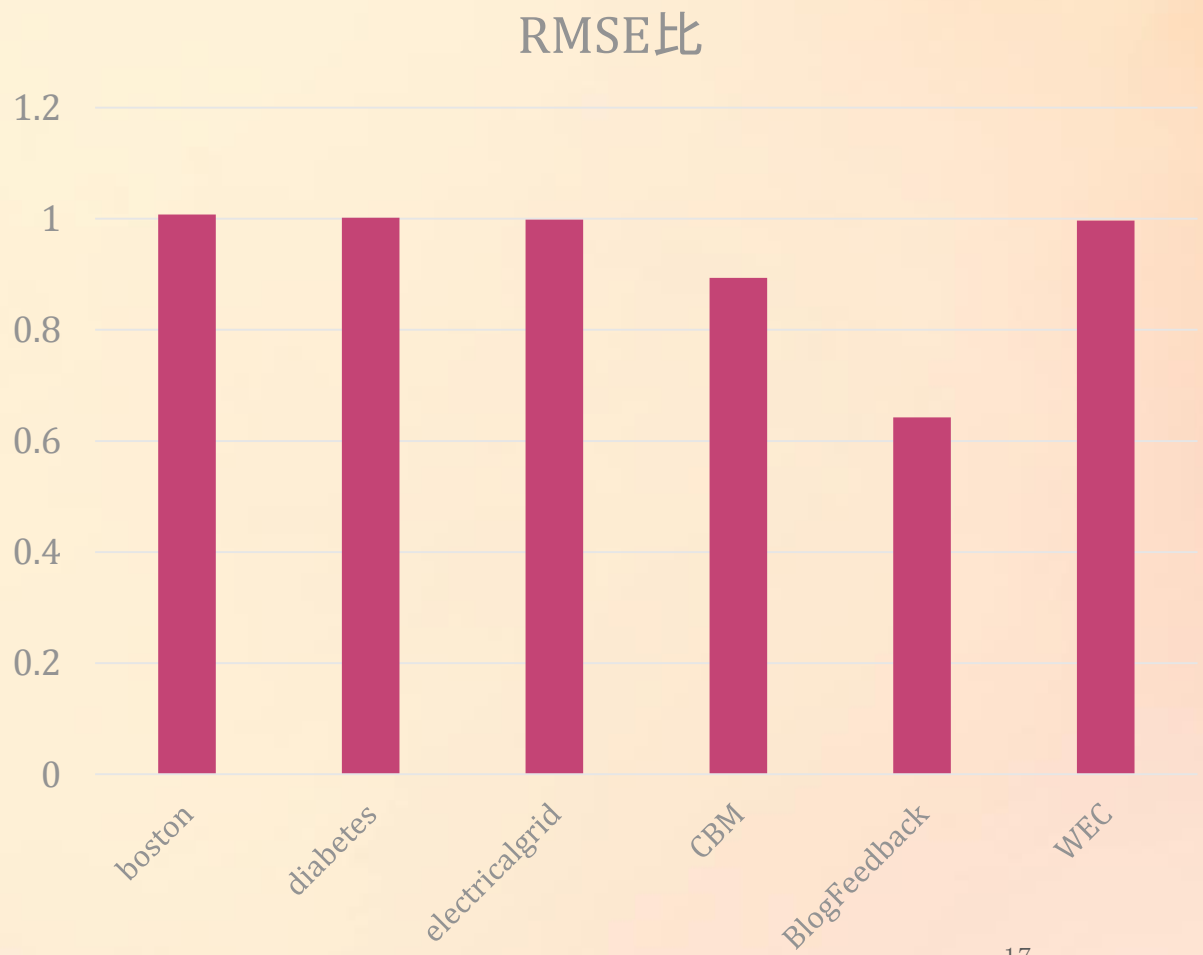
削減率が低い
RNA-Seq PANCAN, arceneなど
→属性の次元が高い，サンプル数が少ない
≒区間集合から求められる共通部分が少ない

回帰問題

データセット名	サンプル数	属性の次元
Boston	506	13
Diabetes	442	10
Electricalgrid	10000	14
CBM	11934	16
BlogFeedback	60021	281
Wave Energy Converter	72000	49

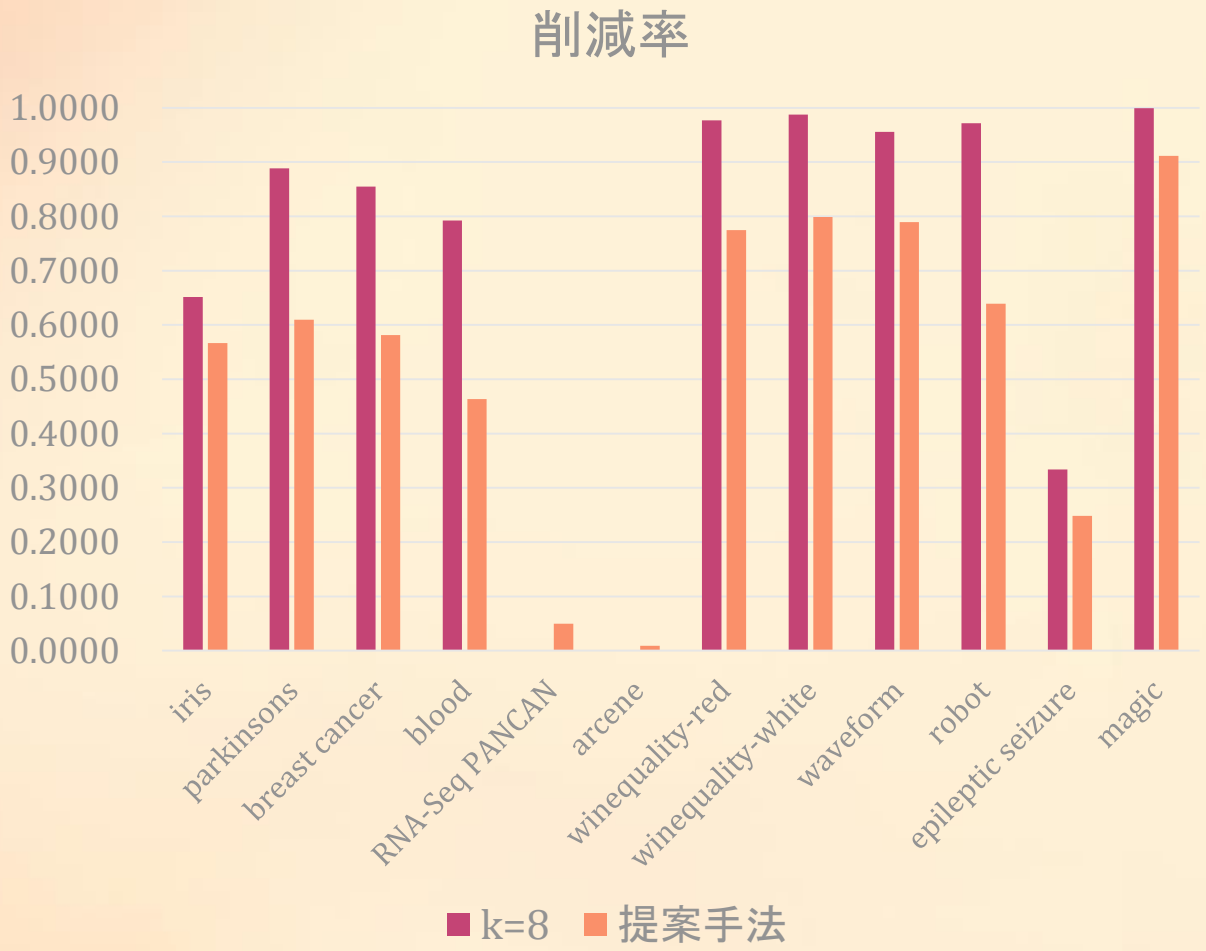
Random Forest 回帰問題 実験結果

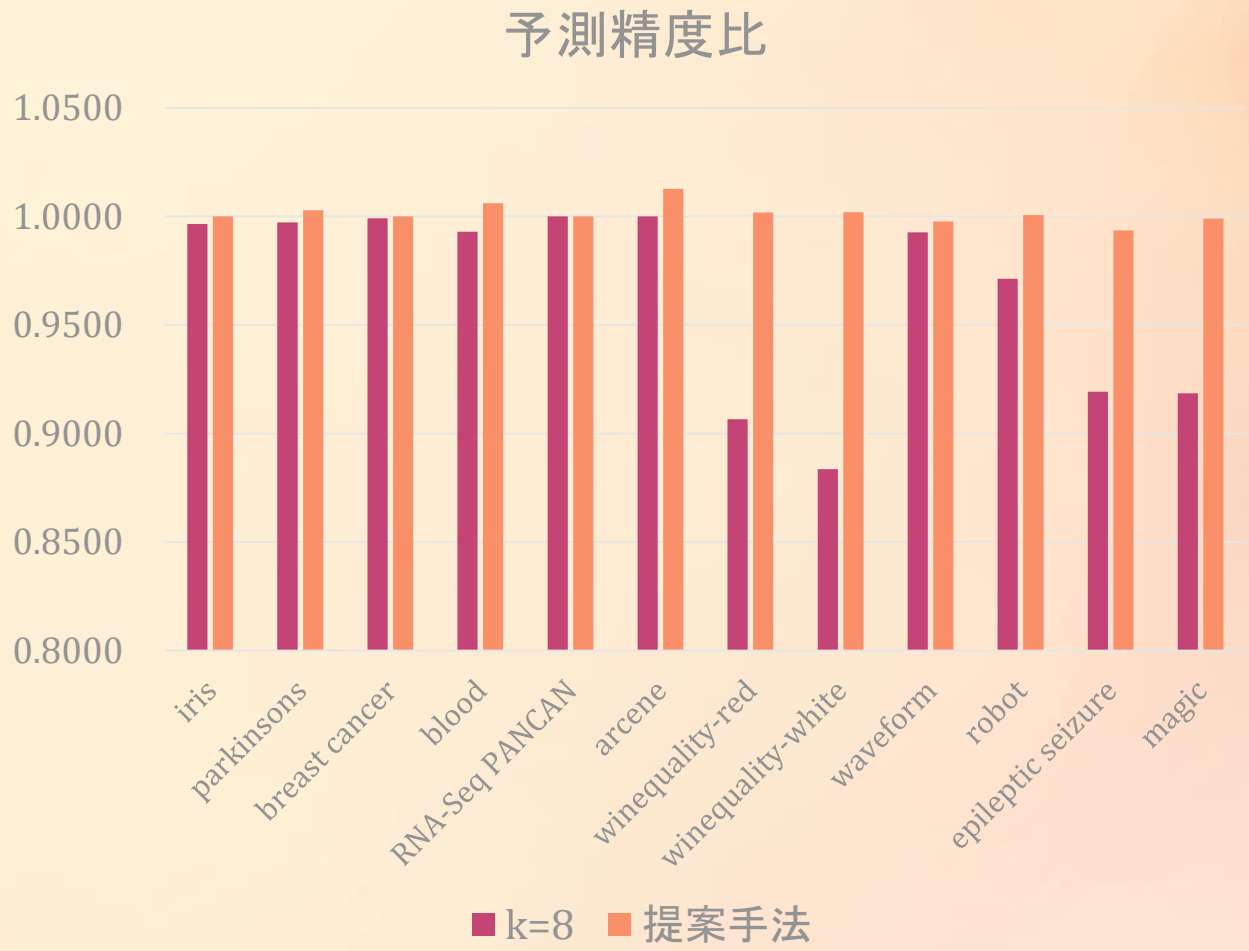




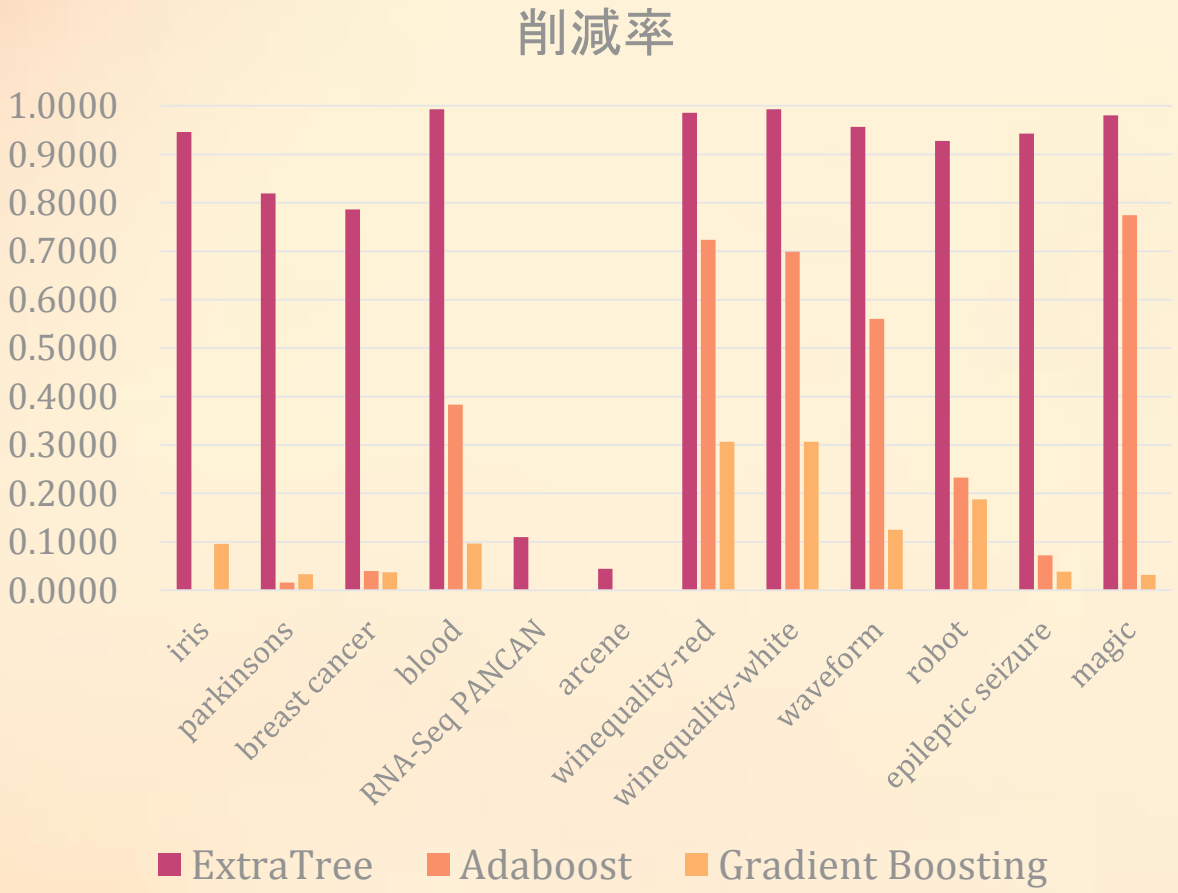
従来手法との比較

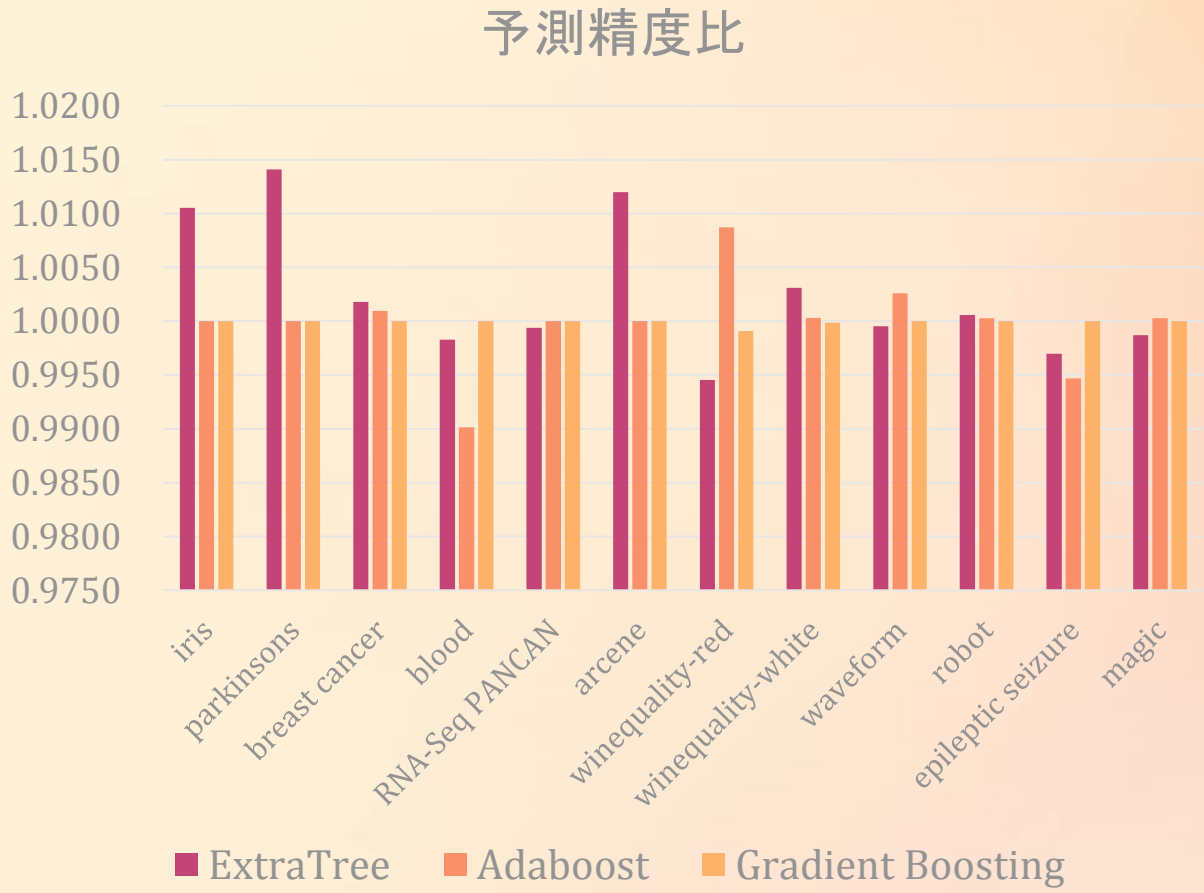
従来手法：k-meansを用いたクラスタリングによる閾値共有化法[Jinguji et al. 2018]





ExtraTrees, Adaboost, Gradient Boosting 実験結果(分類問題)





結論

- 提案手法は予測精度の維持を優先した分岐条件共有アルゴリズムとして実験的に有効
- 分岐条件の削減率が高いのはバギング木だが，ブースティング木でも予測精度に影響はほぼない
- 集積アーキテクチャ研究室(北大)の池田らがハードウェア実装 [*Ikeda, et al., 2020*]
  - 30～50%の計算リソース削減に成功
    →小規模なアンサンブルでは比較器以外の部分のウェイトが高い

学会発表

- ランダムフォレスト識別器の異なる分岐ノードの数の削減
  櫻田 健斗,中村篤祥
  第109回 人工知能基本問題研究会, 2019, pp.62-67
- An Algorithm for Reducing the Number of Distinct Branching Conditions in a Decision Forest
  Atsuyoshi Nakamura and Kento Sakurada
  ECML PKDD, 2019
- 決定木アンサンブル予測器の効率的ハードウェア実装のための簡約化に関する研究
  櫻田健斗, 中村篤祥, 工藤峰一
  第22回 報論的学習理論ワークショップ, 2019