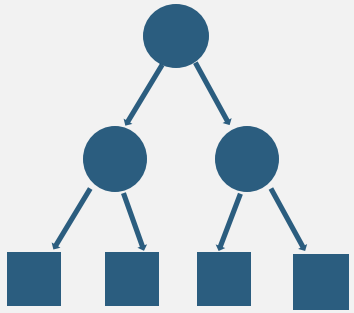
Simplified Decision Tree Induction with Multi-objective Symbiotic Evolution

Otani Laboratory　Takuya Mitarai

決定木と代表的手法

- 決定木・・・木構造を用いた解釈可能な分類・回帰モデル

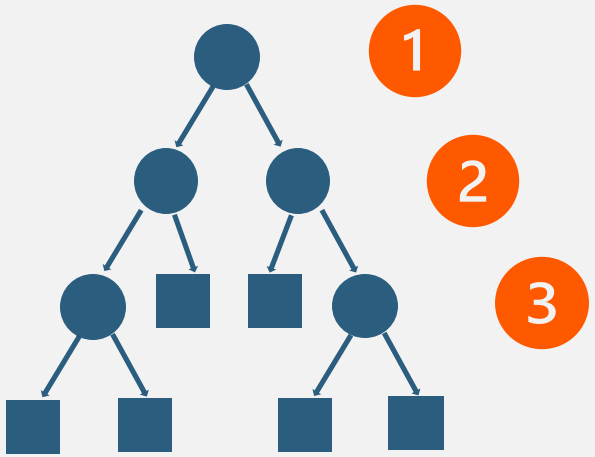
分岐条件が明示的



- 代表的な決定木生成手法

・ID3 [Quinlan, 1986]
・C4.5[Quinlan, 1993]

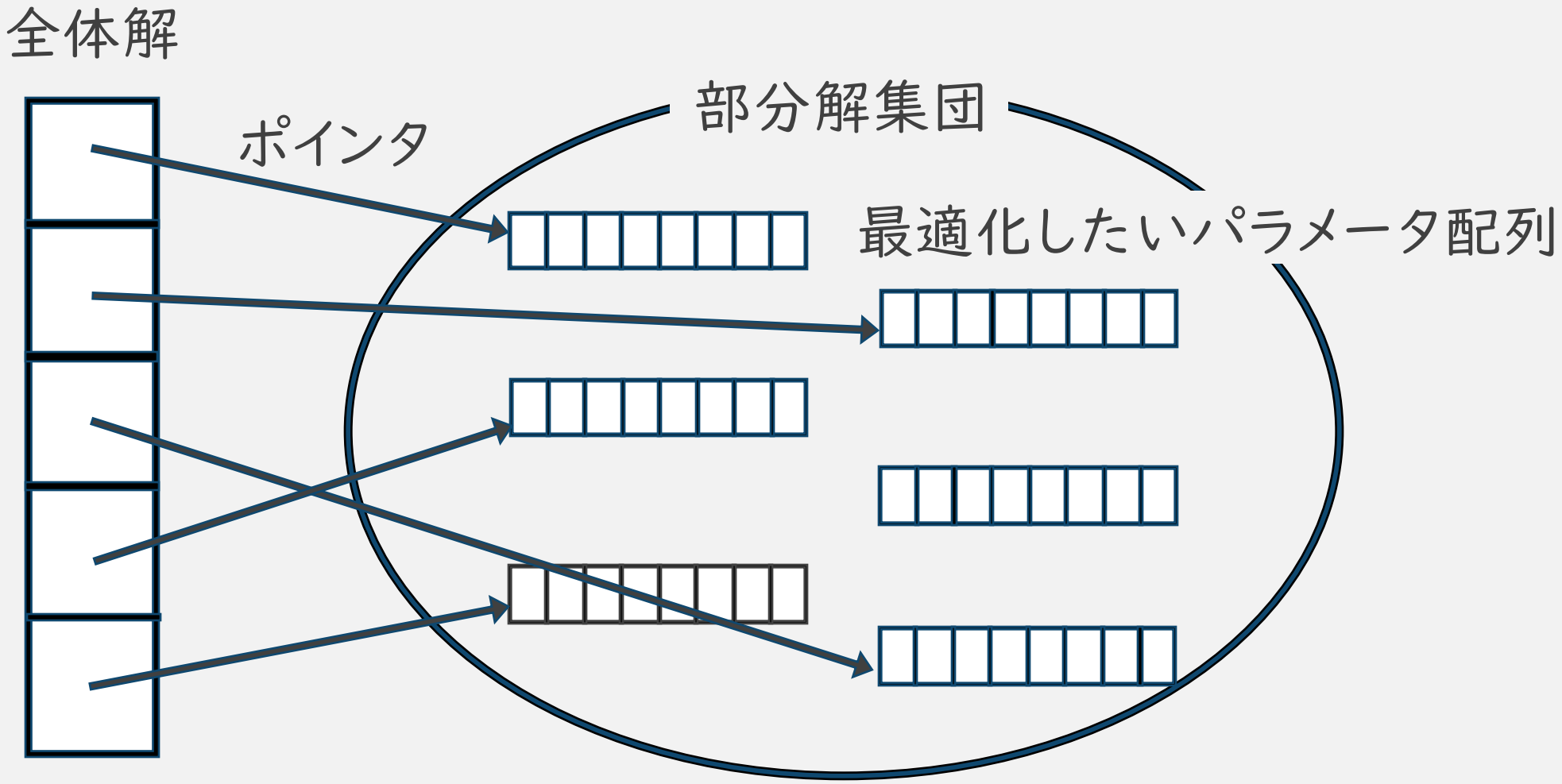
・・・ 貪欲法



✕ ノード数が膨大になり，過学習を起こしやすい

→共生進化を使った決定木生成手法

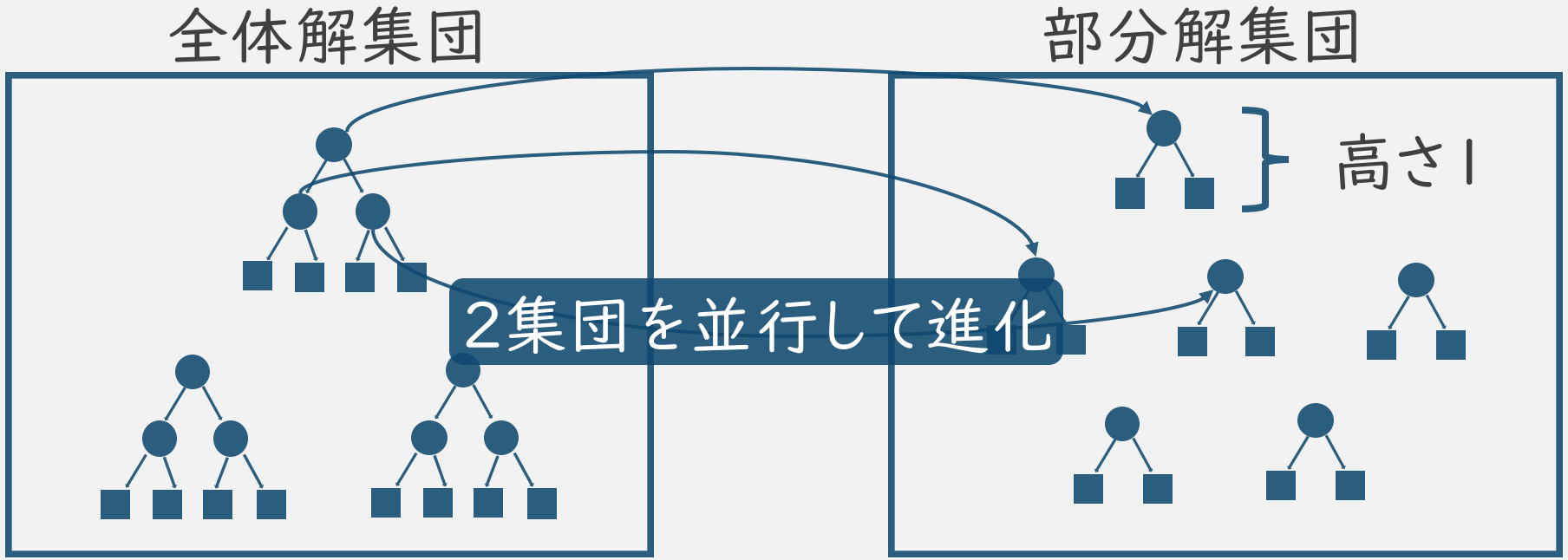
共生進化 ･･･分割統治法に基づく遺伝的アルゴリズム



一般的な遺伝的アルゴリズムよりも探索性能が高い

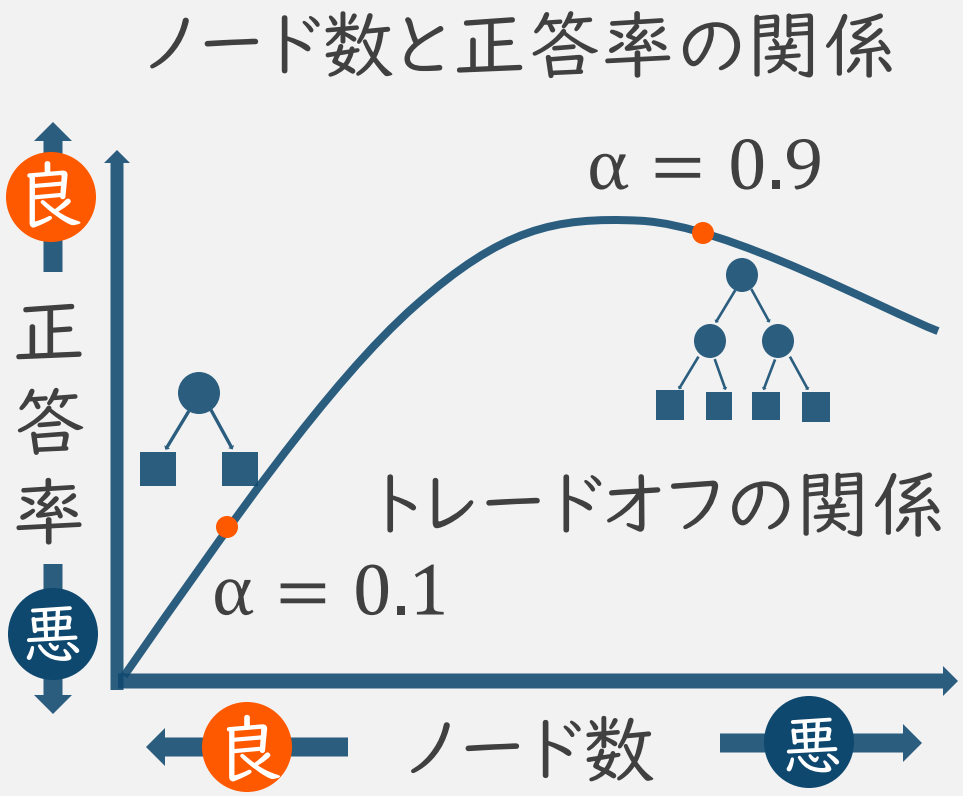
共生進化による簡素な決定木生成手法

- SESAT[大谷, 2004]



○ ノード数を抑えた上で高い正答率を持つ決定木を生成可能

SESATの評価関数



正答率とノード数に関する多目的最適化問題

共生進化は単目的最適化アルゴリズム

$$Fit(\mathrm{X}) = \text{正答率} \times (1 - \alpha \times \text{エントロピー}) \times 100$$

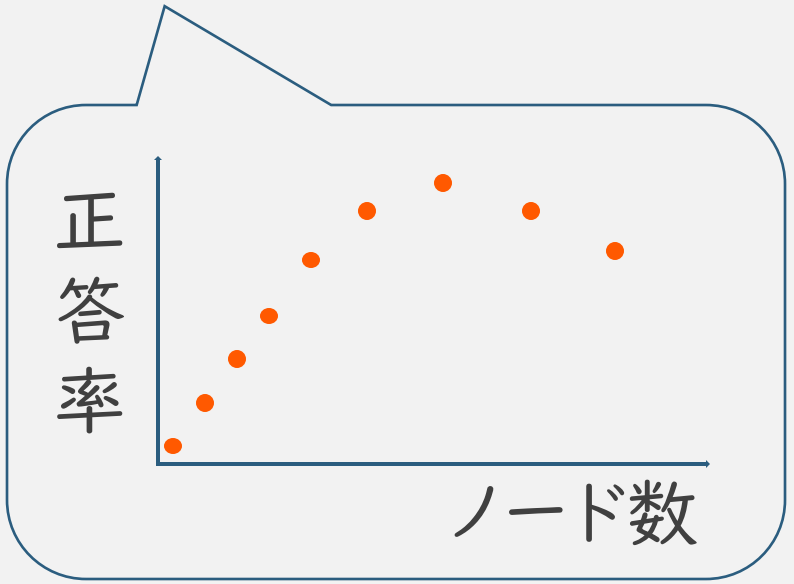
正答率とエントロピーの重み付き積
（ノード数）

✗ 重みαのパラメータ調整の手間

✗ 1回の探索で1つの決定木のみ

研究の目的

🎯 一度の探索で正答率とノード数に関する最適な決定木を列挙



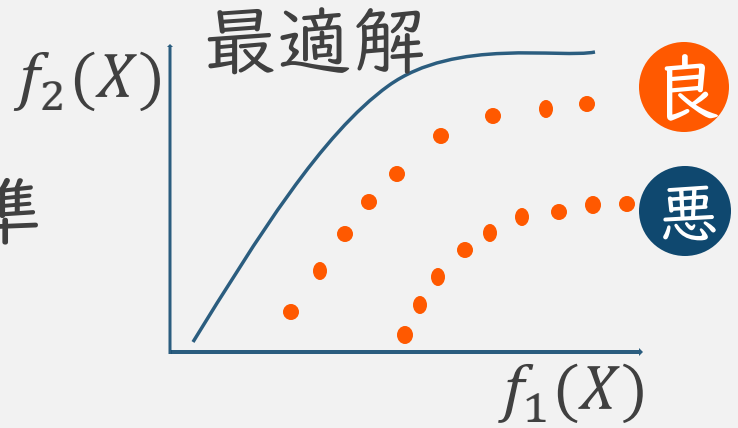
1. 多目的共生進化NSSEを提案

2. NSSESAT(Non-dominated Sorting Symbiotic Evolution for Simple and Accurate Trees)を提案

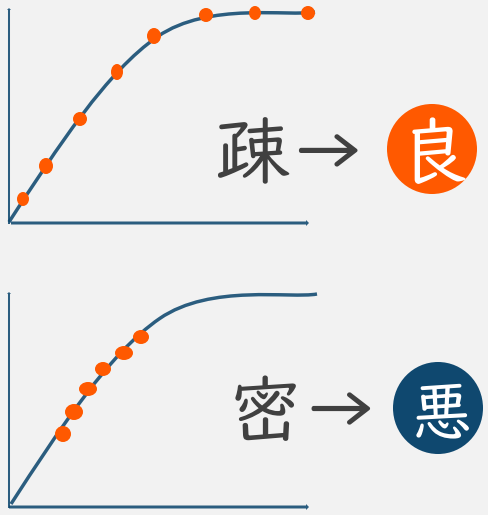
1 多目的共生進化NSSE

- NSGA-II[Deb, 2000]・・・多目的遺伝的アルゴリズム

解の優劣基準



混雑距離



共生進化　単目的最適化

NSGA-IIのエッセンス

NSSE　多目的共生進化

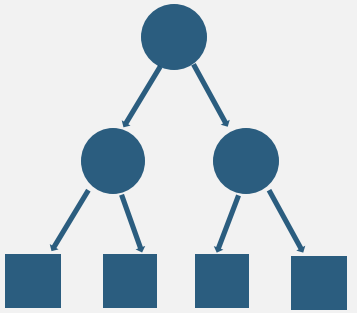
2 NSSESAT<評価関数>

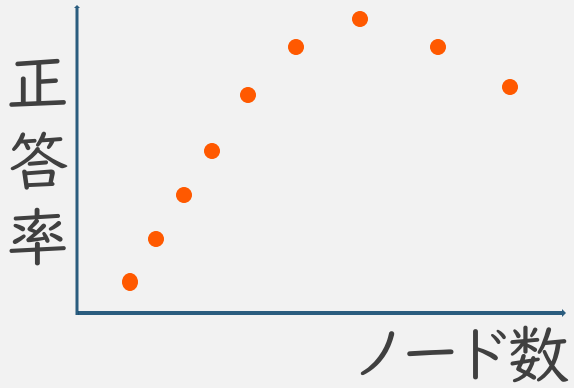
NSSE

解の優劣基準と混雑距離に基づく評価関数

NSSESAT

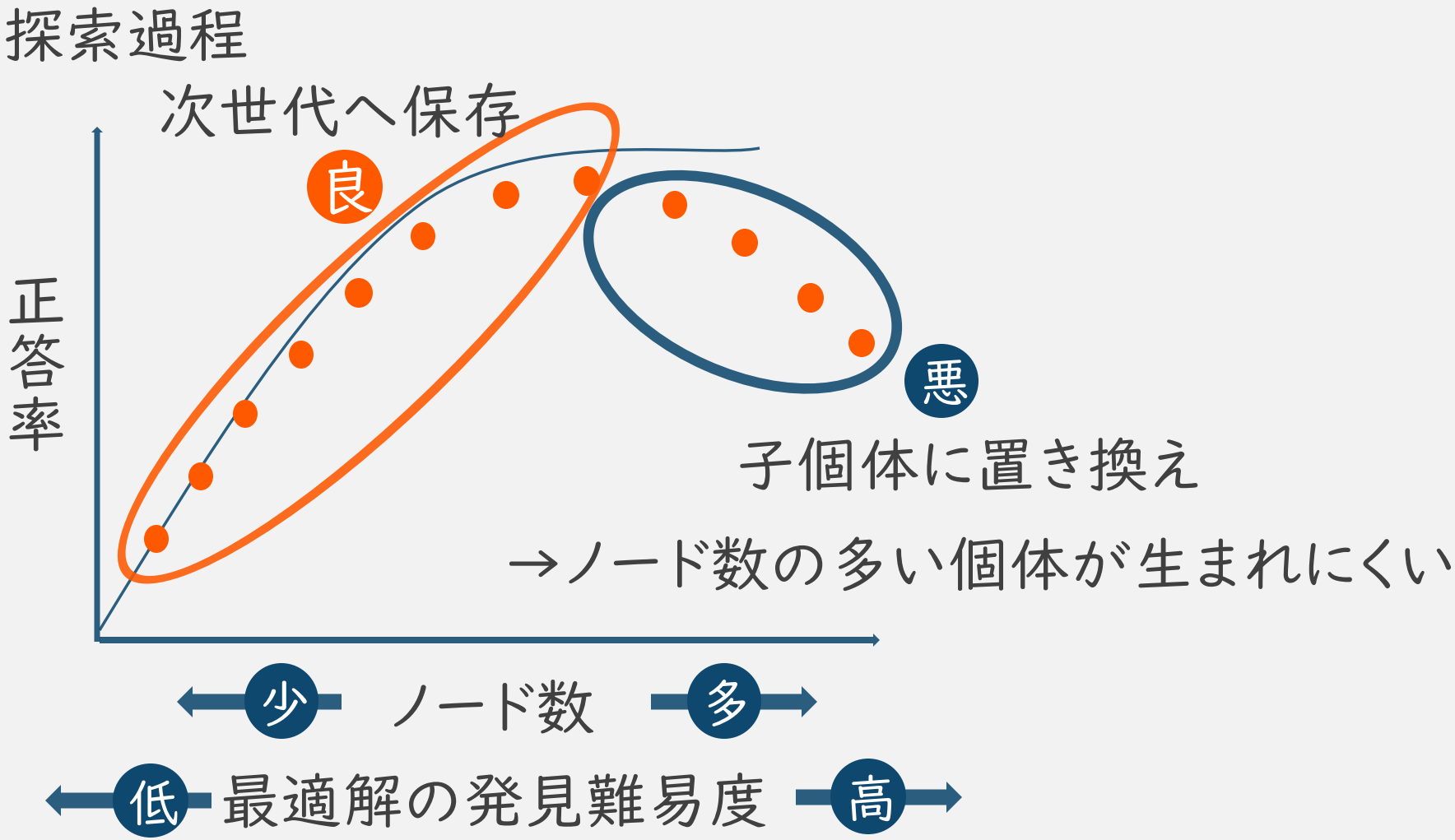
$$Fit(\mathrm{X}) = \text{正答率} \times (1 - \alpha \times \text{エントロピー}) \times 100$$





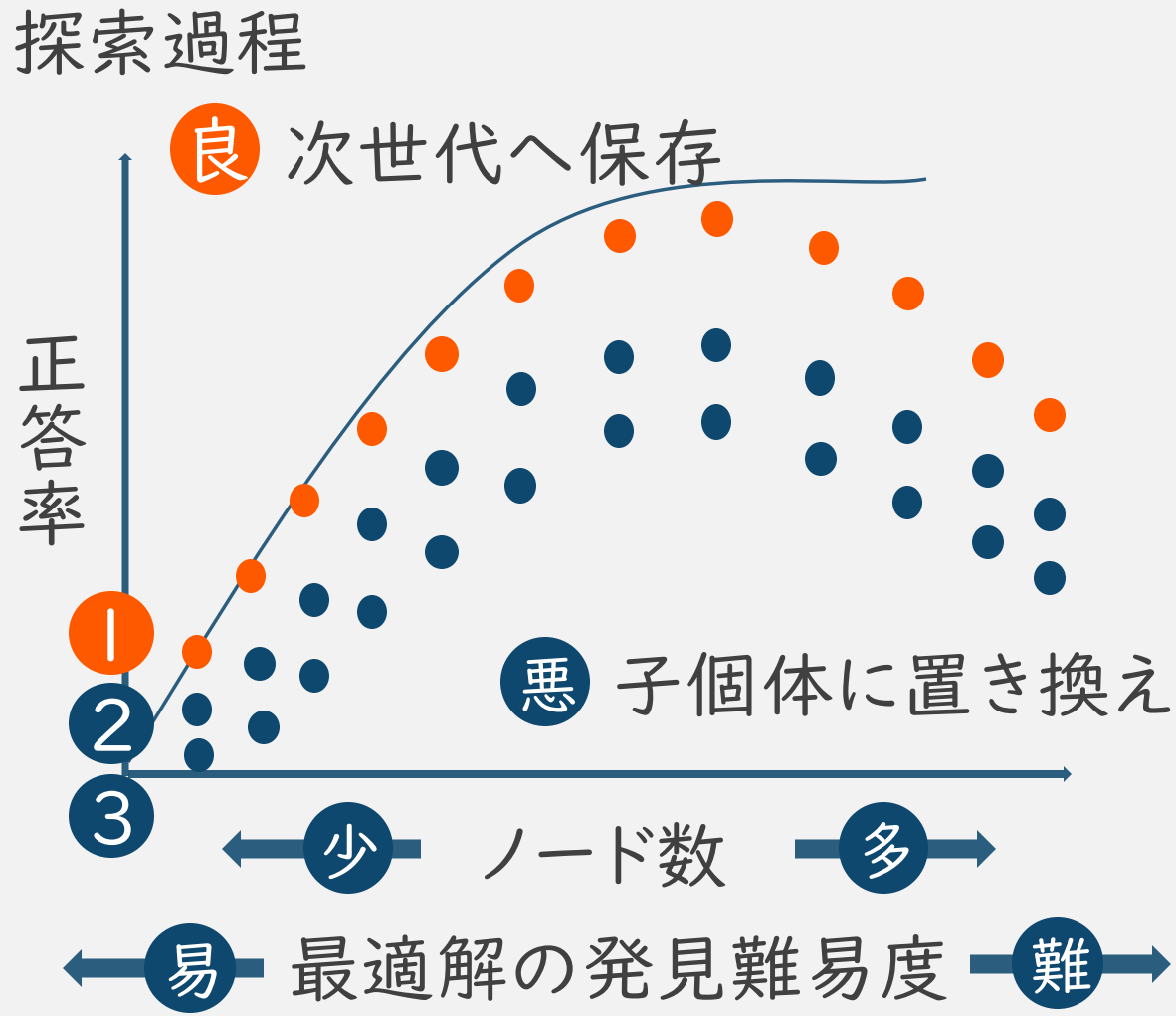
2 NSSESAT<次世代への保存>

NSGA-II



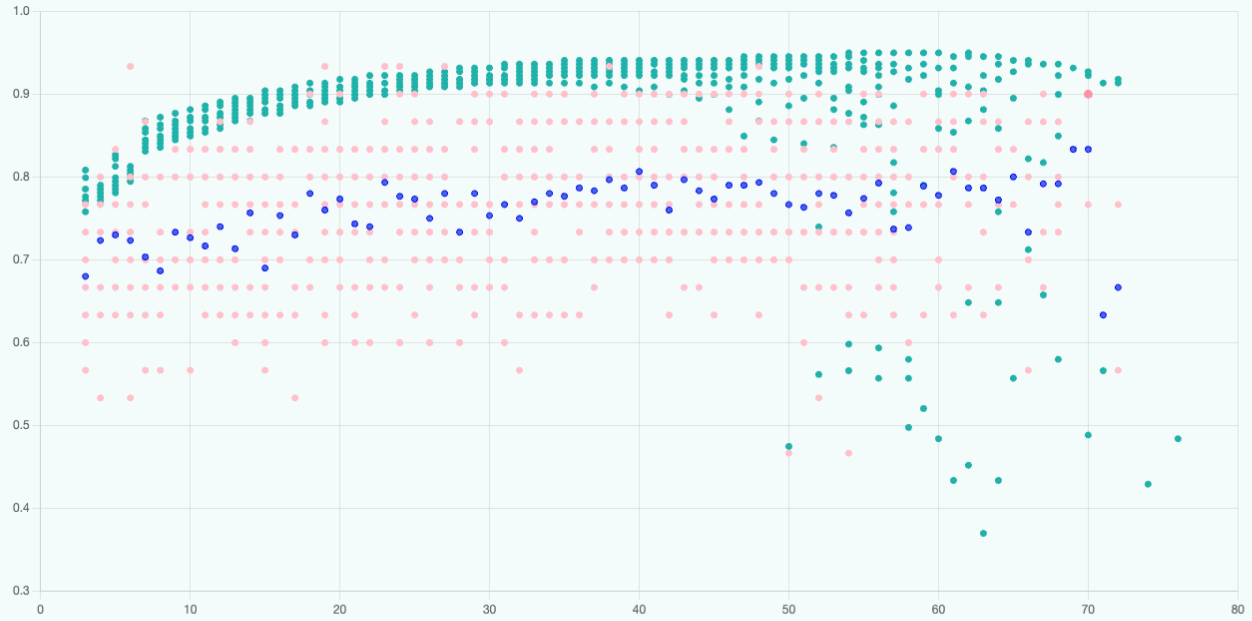
② NSSESAT<次世代への保存>

NSSESAT

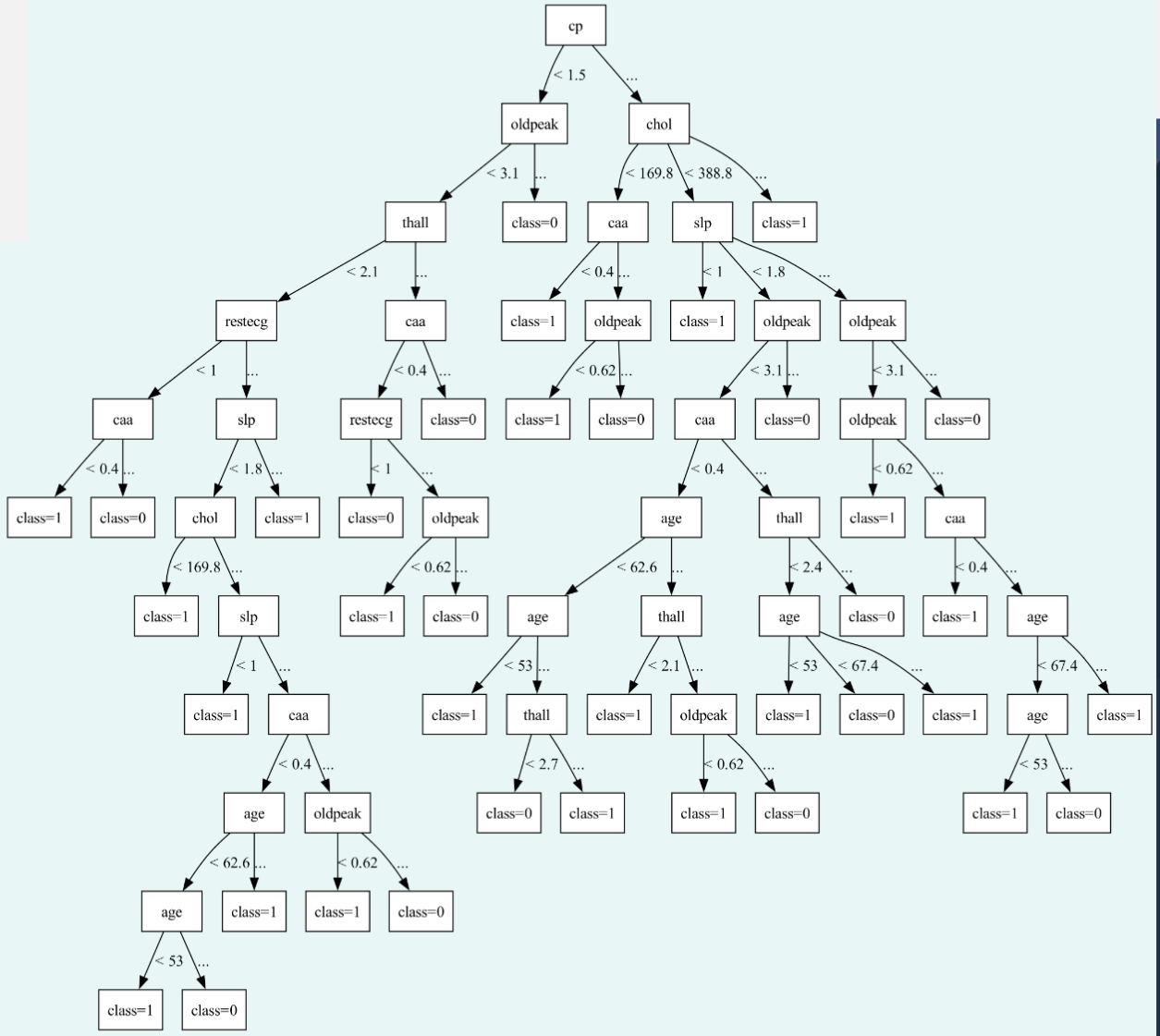


→集団の多様性を向上

どれほど汎化性能があるか確かめたり，
決定木の概形を確認しながら，
モデル選択できるシステム



ノード数	正答率
70	90 %



評価実験 I

- NSSESATとSESATとの比較

UCI機械学習リポジトリのデータセット

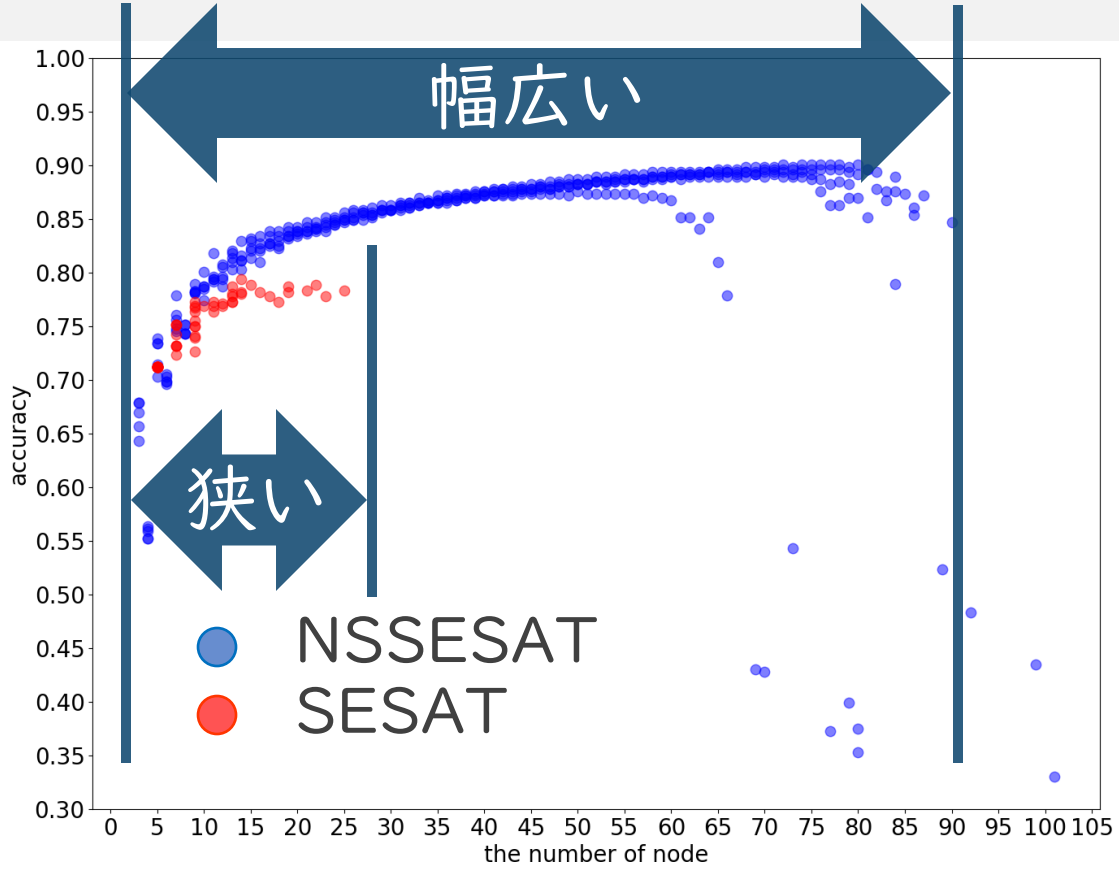
- balance・・・天秤が左か右に傾くか釣り合うか分類する3値問題
- heart・・・心臓発作の確率が高いかどうか分類する2値問題

・SESATでは１１種類の重みαを使って，各5回実行

・ノード数と訓練事例に対する正答率に関する散布図を表示

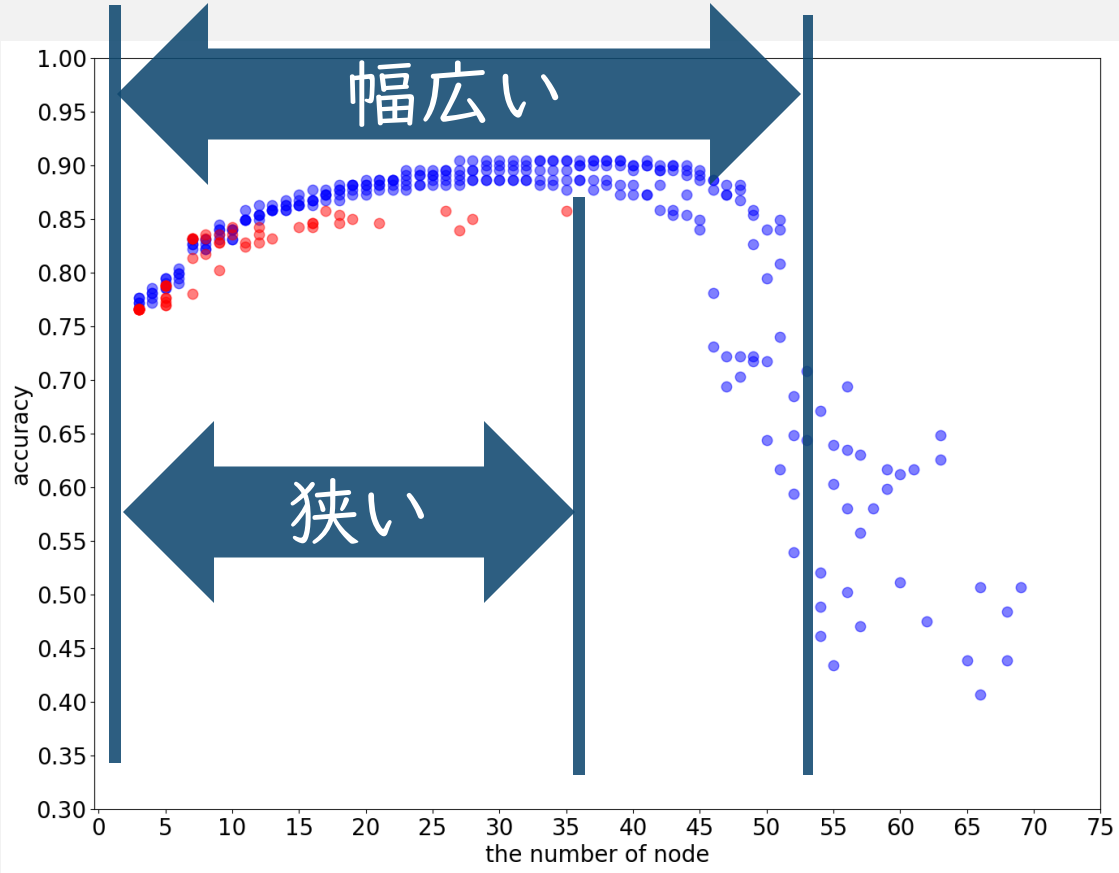
実験結果 I

balance



SESAT ≤ NSSESAT

heart



SESAT ≤ NSSESAT

評価実験2

解釈可能でない

- NSSESATと既存の機械学習手法との比較

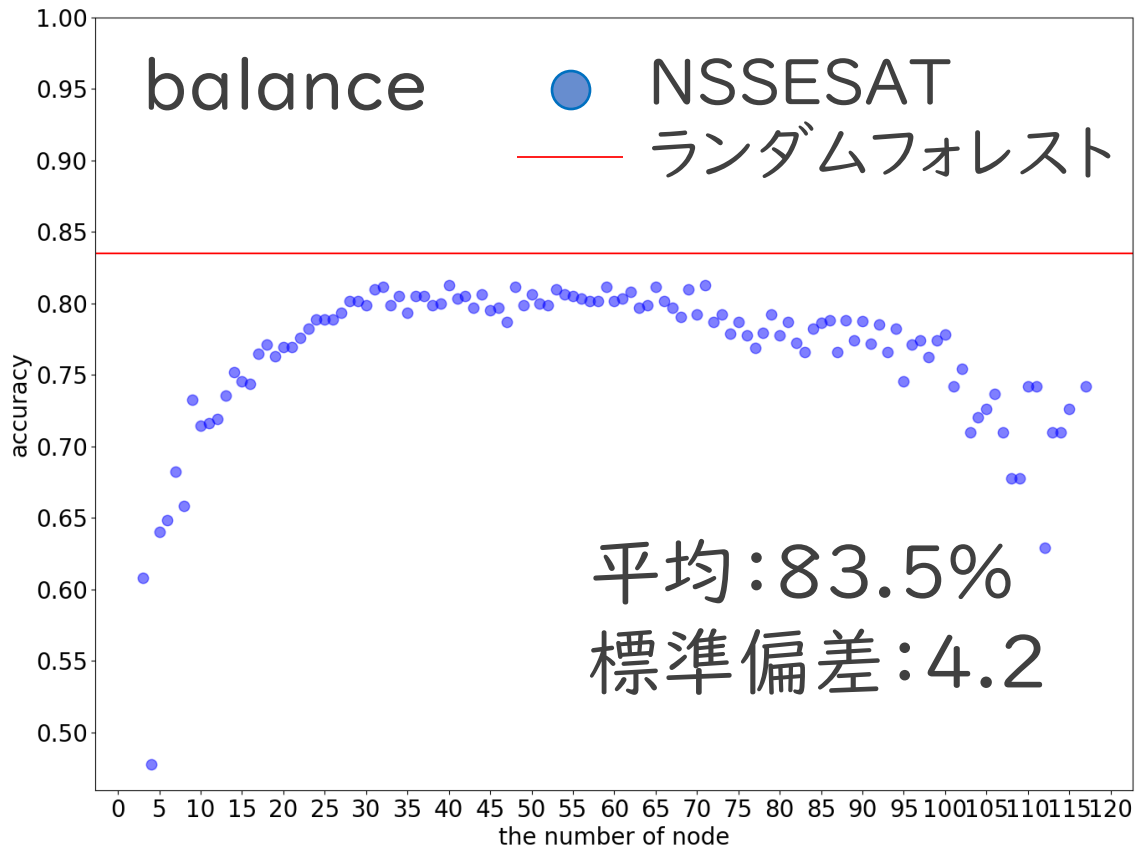
・ランダムフォレスト, RBFカーネルSVM

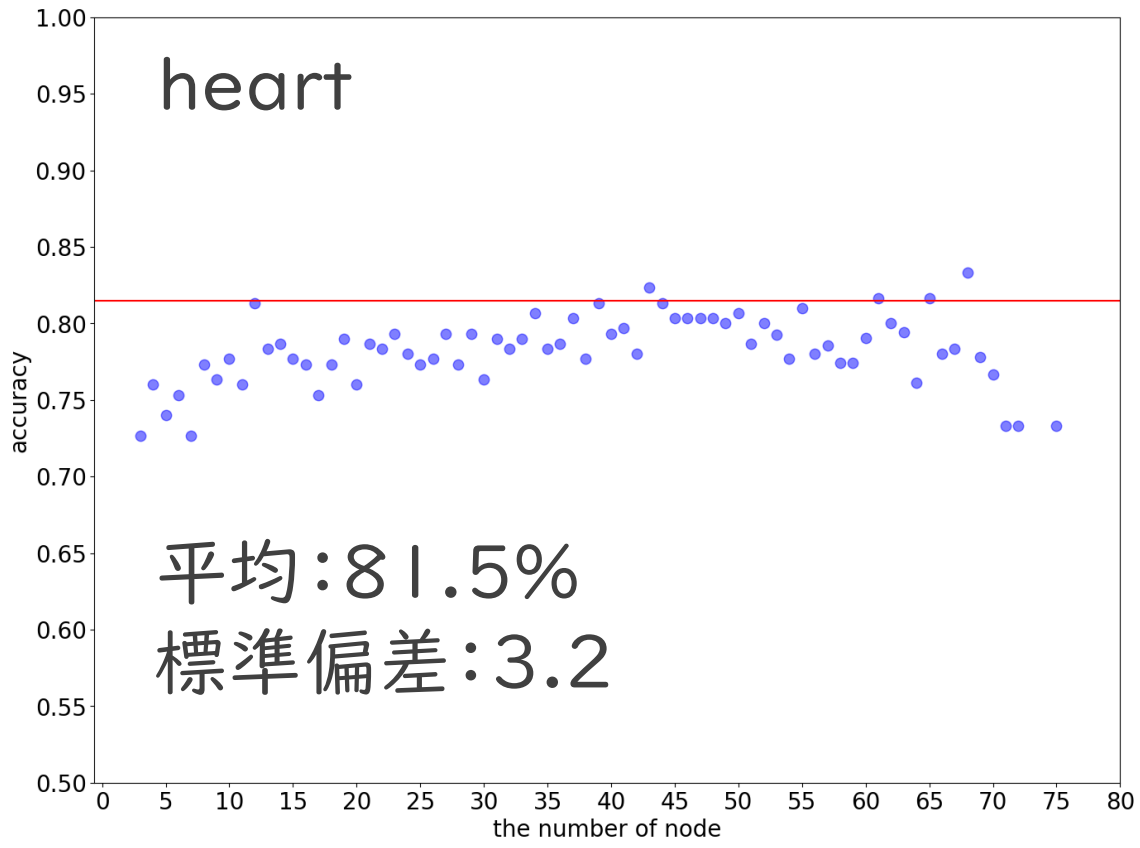
・既存の機械学習手法ではネストした交差検証を行う

・10-fold交差検証を行い, 正答率の平均値をプロット

・評価実験1と同じデータセットを用いる

実験結果2　vs ランダムフォレスト

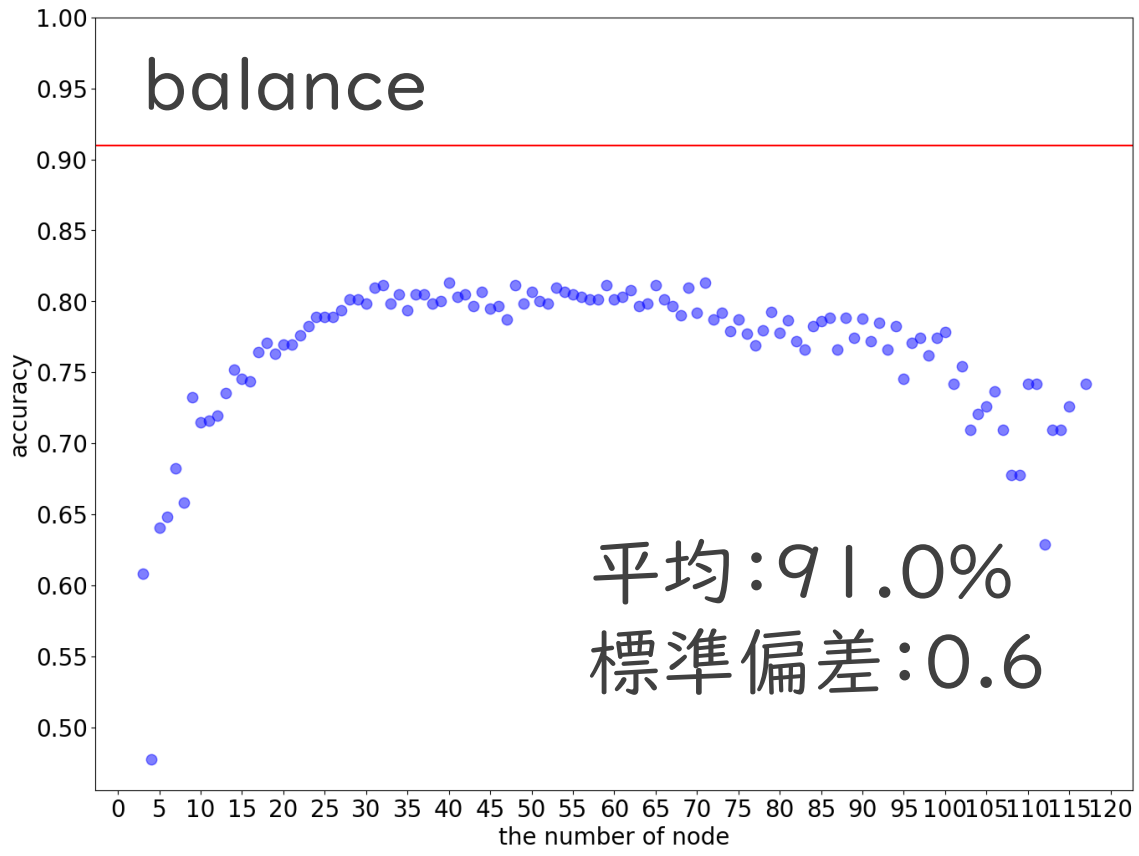


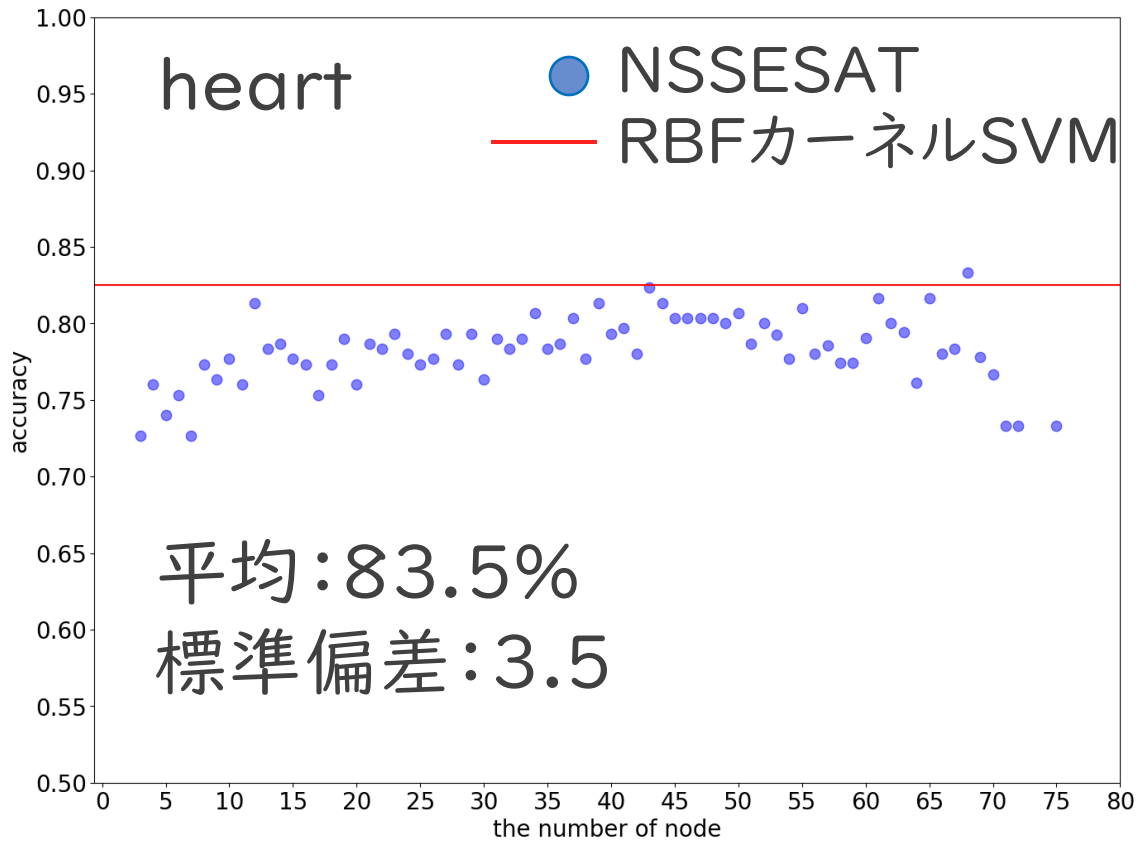


NSSESAT ≤ ランダムフォレスト

NSSESAT ≈ ランダムフォレスト

実験結果2　vs RBFカーネルSVM





NSSESAT ≤ RBFカーネルSVM

NSSESAT ≈ RBFカーネルSVM

考察

- ノード数と正答率に関する最適な決定木を一度に列挙可能
- 一部の問題に対して，既存の機械学習手法と同等の性能を持つ

解釈可能な分類モデルを生成したい

既存手法と同等の正答率を維持したい

→ NSSESAT