ランダムフォレスト
Random Forest
RF

明治大学 理工学部 応用化学科
データ化学工学研究室　金子 弘昌

Random Forest (RF) とは？

✓サンプルと説明変数とをランダムにサンプリングして、
決定木をたくさん作る

✓複数の決定木の推定結果を統合して、最終的な推定値とする

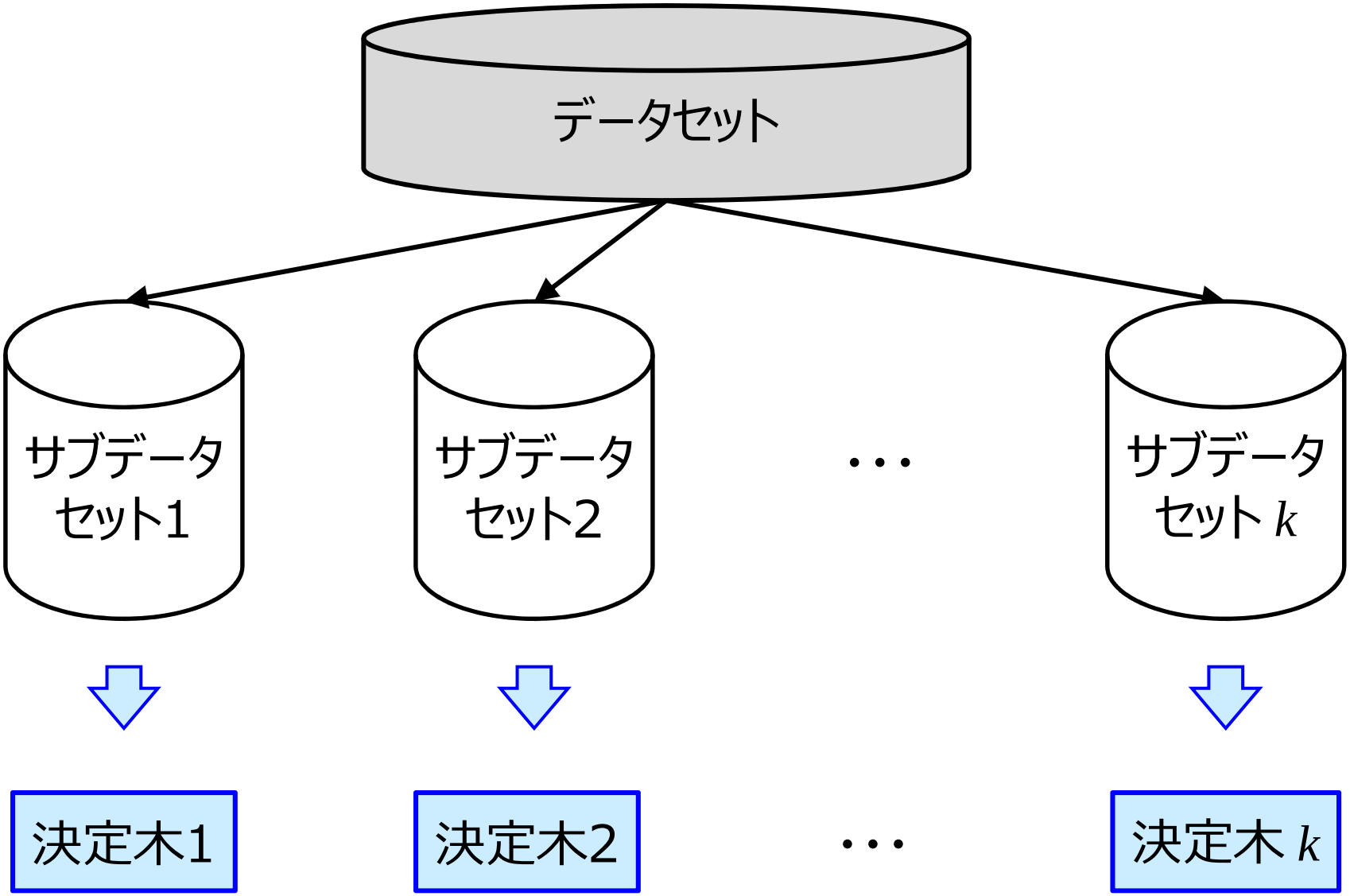
✓アンサンブル(集団)学習 (Ensemble learning) の１つ

✓決定木と比べて精度は高くなることが多いが、モデルを解釈することは
難しい

✓回帰分析にもクラス分類にも使える

✓説明変数の重要度を議論できる

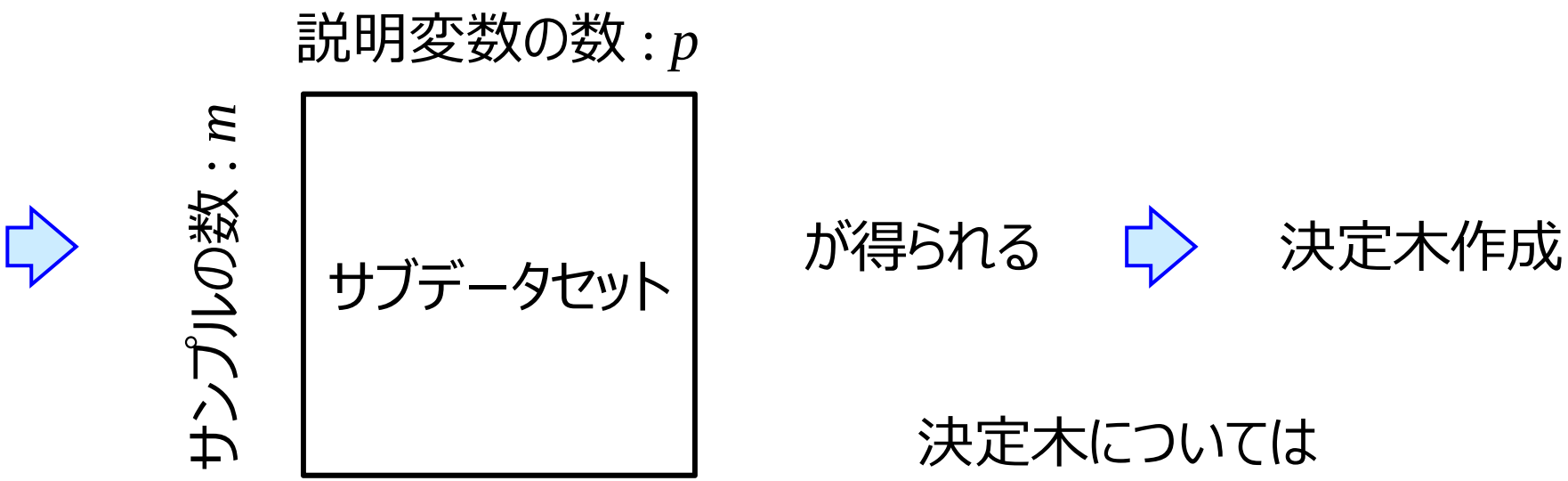
RFの概略図



どのようにサブデータセットを作るか？

✓データセットのサンプル数: m
- サンプルを重複を許してランダムに m 個選択

✓データセットの説明変数(記述子)の数: n
- 説明変数を重複を許さずランダムに p 個選択



決定木については
http://datachemeng.com/decisiontree/

サブデータセットの数・説明変数の数はどうする？[4]

✓グリッドサーチ ＋ クロスバリデーション

✓サブデータセットの数の候補 例

- 100, 200, 300, 400, 500

✓説明変数の数の候補 例

- データセットにおける説明変数(記述子)の数の
  10, 20, …, 80, 90 %

どのように推定結果を統合するか？

✓回帰分析

- k 個の推定値の平均値

✓クラス分類

- k 個のクラス分類結果で多数決

説明変数 (記述子) の重要度

✓説明変数 (記述子) の重要度 I_j

$$I_j = \frac{1}{k} \sum_{T} \sum_{t \in T, j} \frac{m_t}{m} \Delta E_t$$

k : サブデータセットの数 (決定木の数)
m : サンプル数
T : ある決定木
t : T におけるあるノード
ΔE_t : t にしたときの E (決定木における評価関数) の変化

- ✓ 目的変数の誤差の二乗和 (回帰分析)
- ✓ Gini 係数など (クラス分類)

＊ Scikit-learn では変数の重要度としてこれを採用

Out-Of-Bag (OOB)

✓サブデータセットを作るとき、m 個のサンプルから重複を許して m 個のサンプルを選択

⇨ サブデータセットごとに、選ばれなかったサンプル (Out-Of-Bag, OOB) が存在

⇨ OOBにより、各決定木の予測性能を検討可能

OOBを用いた説明変数（記述子）の重要度

✓説明変数 (記述子) の重要度 I_j

$$I_j = \frac{1}{k}\sum_{i=1}^{k}\left(F_i - E_i\right)p\left(j\right)$$

k : サブデータセットの数 (決定木の数)

$p(j)$: i 番目の決定木に変数 j が使われていたら 1, そうでなければ 0

E_i : i 番目の決定木において、OOBを推定したときの

- ✓ 平均二乗誤差 (回帰分析)
- ✓ 誤分類率 (クラス分類)

F_i : i 番目の決定木を作成した後に、説明変数をランダムに並び替えて、OOBを推定したときの

- ✓ 平均二乗誤差 (回帰分析)
- ✓ 誤分類率 (クラス分類)

E_i が小さいほど、F_i が大きいほど、I_j が大きい
→ j 番目の説明変数 (記述子) の重要度が高い

決定木の設定はどうする？

✓各決定木の深さはどのように設定すれば良い？

- とりあえず深くすればよく、最適化する必要はない

- なぜ？

- モデルの誤差は バイアス＋バリアンス で表現できる
- ランダムフォレストのようなアンサンブル学習におけるバギングの場合、バリアンスは減らせてもバイアスは減らせない
- 決定木を過学習させて、(バリアンスは大きくなってしまう一方で)バイアスを小さくしておく
- アンサンブル学習により、大きくなってしまったバリアンスを小さくする