ランダムフォレストによる
クラス分類

平成29年 12月 1日

富山県立大学　杉山桃香

発表内容

- 背景・目的
- ランダムフォレストの概要
- ランダムフォレストの仕組み
- 従来の手法とランダムフォレストの比較
- まとめ

背景・目的

- 背景

→ 今、分類や回帰の問題を扱う場合選択する手法が多種多様

→ ランダムフォレストに従来手法と比べてどんな利点があるか

- 目的

→ ランダムフォレストの概要の確認

→ ランダムフォレストの仕組みを理解し特徴を理解する

→ 従来の手法と比べたランダムフォレストの利点について考える

ランダムフォレストとは?

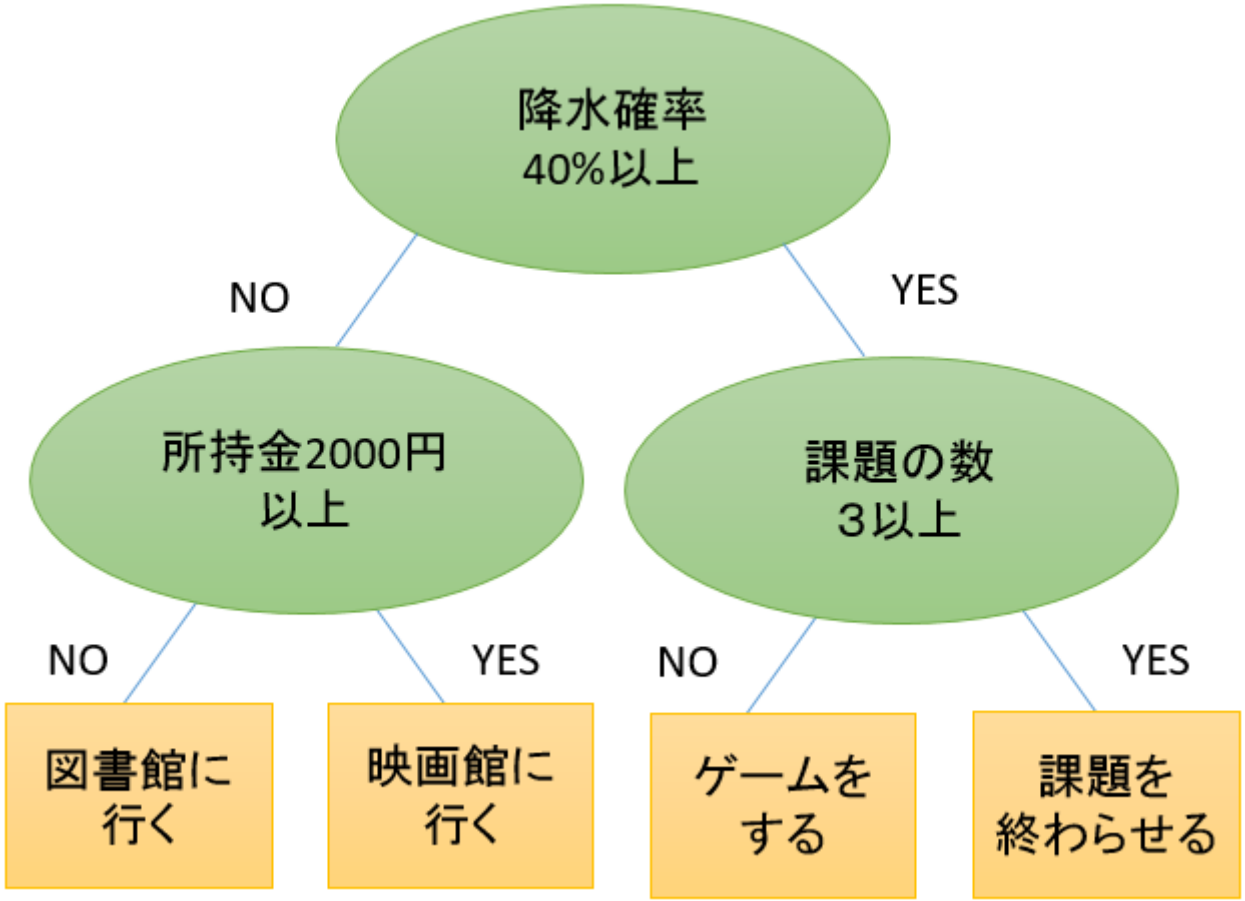
- 高精度の機械学習アルゴリズム
- **決定木**を用いた**集団学習**の一つである
- 識別関数に分類される
- 教師データを必要とする

ランダムフォレストの仕組み

1. 決定木の仕組み
2. どうして集団学習するの?
3. 他の集団学習と違うの?
4. 説明変数の重要度について

1.決定木の仕組み

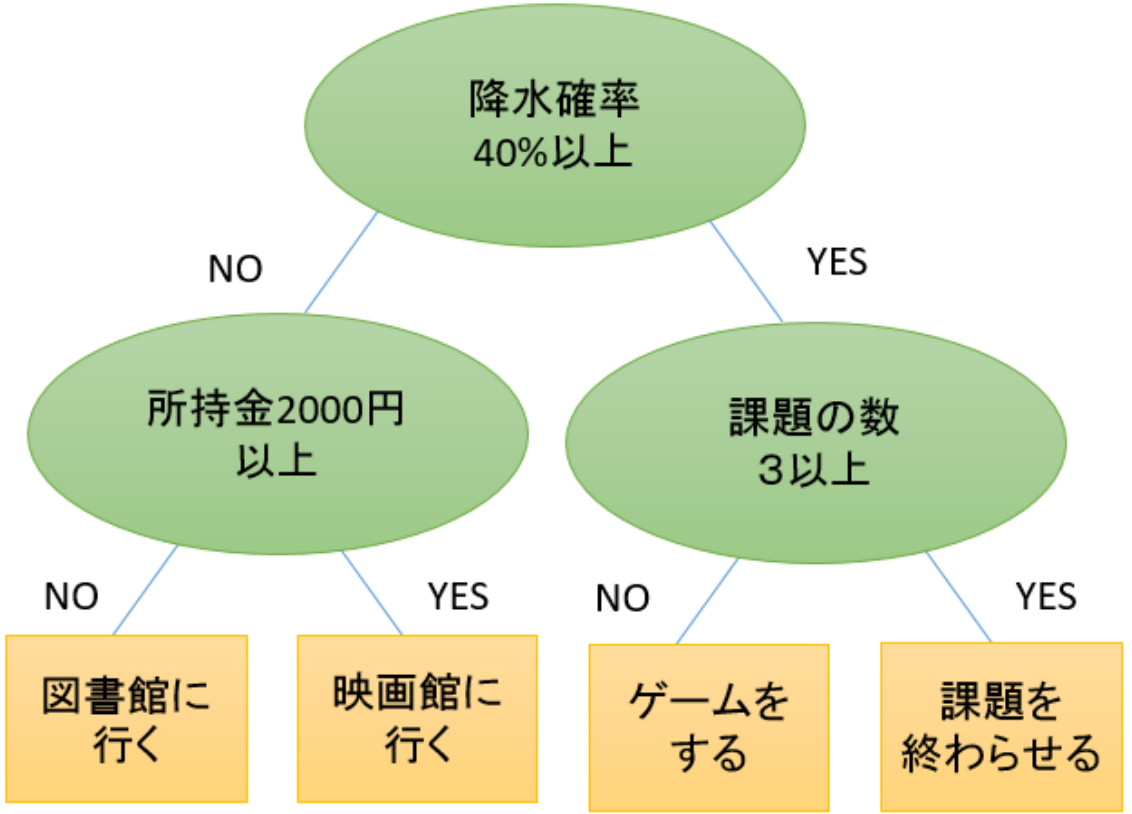
どうやって決定木を作る??



そもそも決定木とは?

ツリー構造のグラフを作成し、予測/分類を行う機械学習のアルゴリズム

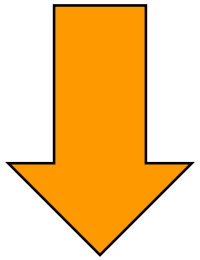
弱学習機（精度の低い学習器）に分類される



決定木を作るポイント

質問はたくさん情報を引き出せる順に並べる

どうやってたくさん情報を引き出せる質問を判断するのか?



引き出せる情報量を数値化する

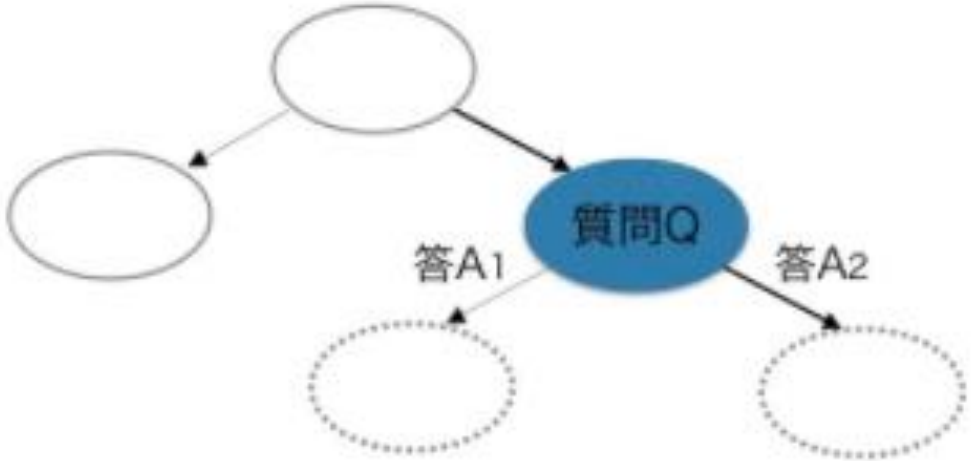
数値化するとは??

- ある質問得る情報量を**情報ゲイン**と呼ぶ。
- **情報ゲイン**は次式で表される。

$$\Delta I = P(Q)I(Q) - \sum_i P(A_i)I(A_i)$$

$P(Q)$: 前のノードからノード Q にくる確率

$I(Q)$: ノード Q におけるエントロピー(関数)



$$\Delta I = P(Q)I(Q) - \sum_i P(A_i)I(A_i)$$

- $I(A)$ のバリエーション

※ $P(k|A)$: ノードAで選択肢kが選ばれる確率

- ①エントロピー(系の取りうる状態数の指標)

$$I(Q) = -\sum_A P(A|Q)\log P(A|Q)$$

→ 高い：乱雑な状態

→ 低い：整理整頓された状態

- ②Gini係数(系の不純度の指標)

色んな目がごっちゃにでるサイコロ

$$I(Q) = 1 - \sum_A \left(P(A|Q)\right)^2$$

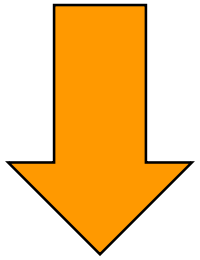
→ 高い：純度が低い状態

→ 低い：純度が高い状態

↑大きいと純度が高い(2回続けて同じ目がでるサイコロは純度が高い！)

$$\Delta I = P(Q)I(Q) - \sum_i P(A_i)I(A_i)$$

- どちらにせよ、エントロピー関数またはジニ係数$I(A)$が低い値を示すと、情報ゲインが増える



情報ゲインを最大にする質問を、各ノードで見つけ決定木を作る

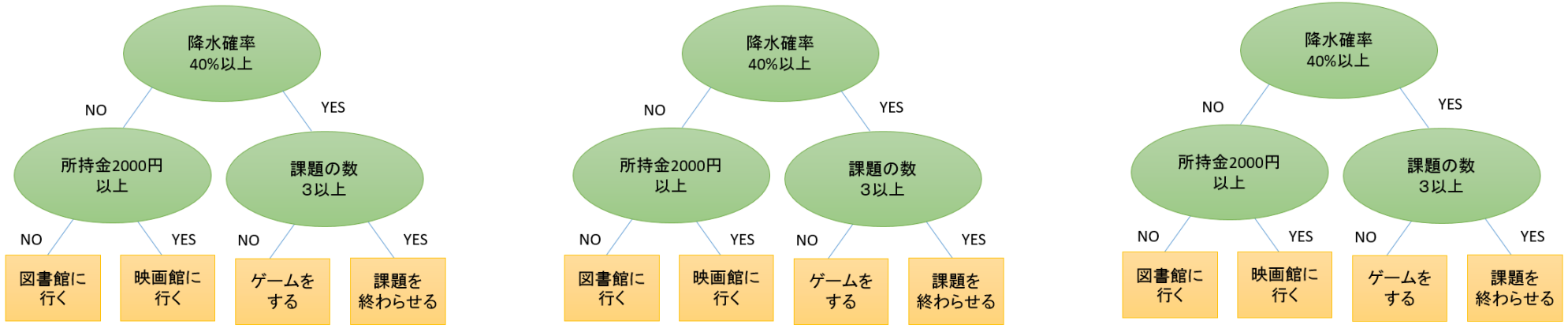
2.どうして集団学習するの?

ランダムフォレストに限らず、弱学習器では集団学習が有効な場合が多い

どんなケースで集団学習が有効になるのか?

集団学習とは?

集団学習は、アンサンブル学習とも呼ばれる



結果をまとめ合わせて１つの識別器を構築する学習方法

汎化誤差について考える

分類/回帰のいずれにおいても、**汎化誤差**(新たなデータに対する誤認識率)は

$$\text{汎化誤差} = \text{バイアス} + \text{バリアンス} + \sigma$$

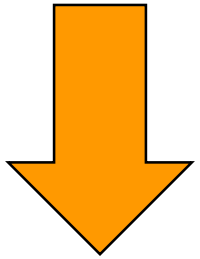
※バイアス (学習モデルの単純さに由来する誤差)
※バリアンス(学習データの違いに由来する誤差)
※σ(削除不能な誤差)

■線形の様な**単純なモデル**では、**バイアスが大きくバリアンスが小さい**
単純なモデルなので、ノイズに強いが複雑な表現はできない。
(例)SVM、最小二乗法など

■高次の**複雑なモデル**では、**バイアスは小さいがバリアンスは大きい**
複雑な表現が可能だが、過学習してしまいがち。ノイズも再現してしまう。
(例)ニューラルネット、決定木など

なんで集団学習するの?

集団学習は**バリアンスを下げる**為に用いられる



複雑な表現が可能だが、過学習してしまいがちなアルゴリズムに向いている

3.他の集団学習との違いは

- **Bagging**

  全教師入力データからランダムにとったデータで、複数の学習器を作成する。

- **Boosting**

  説明変数の重要度を逐次的に変更し、複数の学習器を作成する。

- **Random Forest**

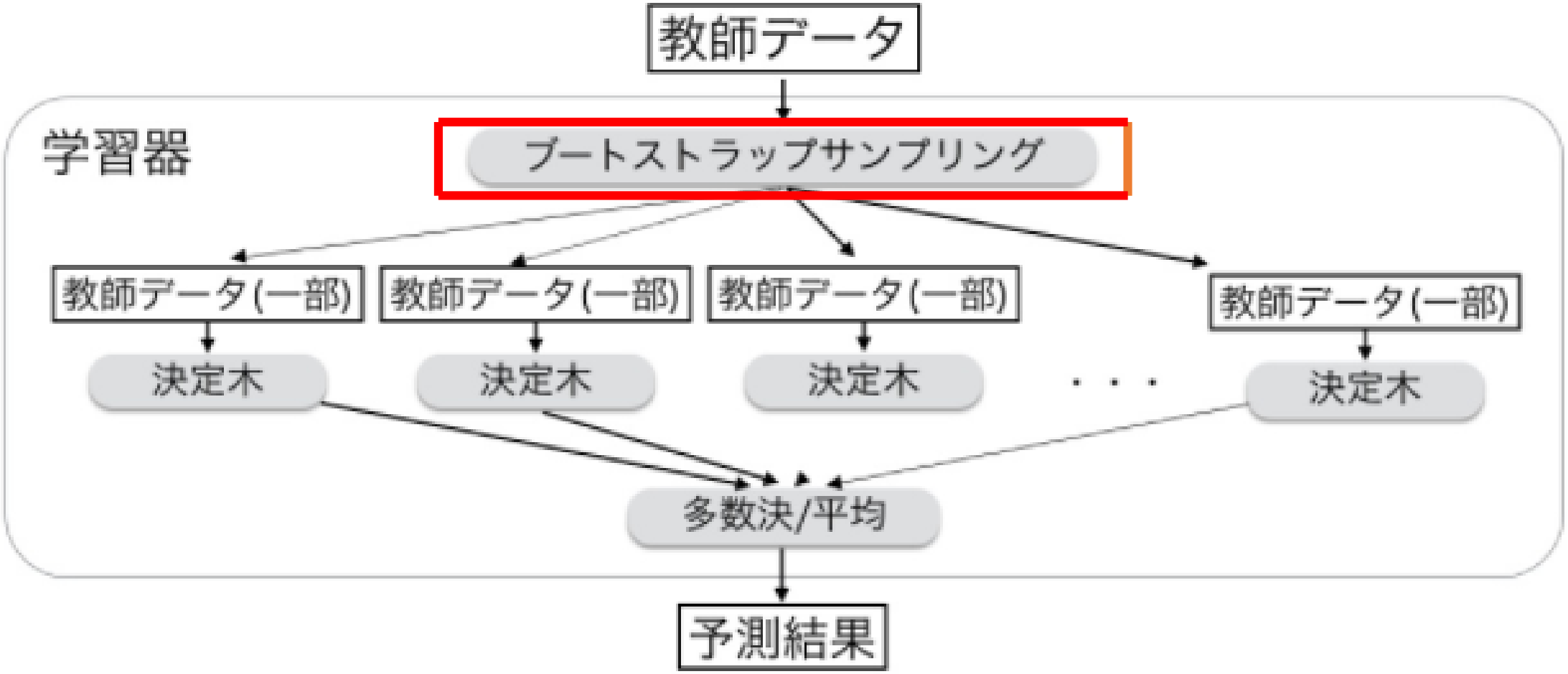
  決定木を学習とする、集団アルゴリズム

  説明変数もランダムにする点がBaggingと異なる。

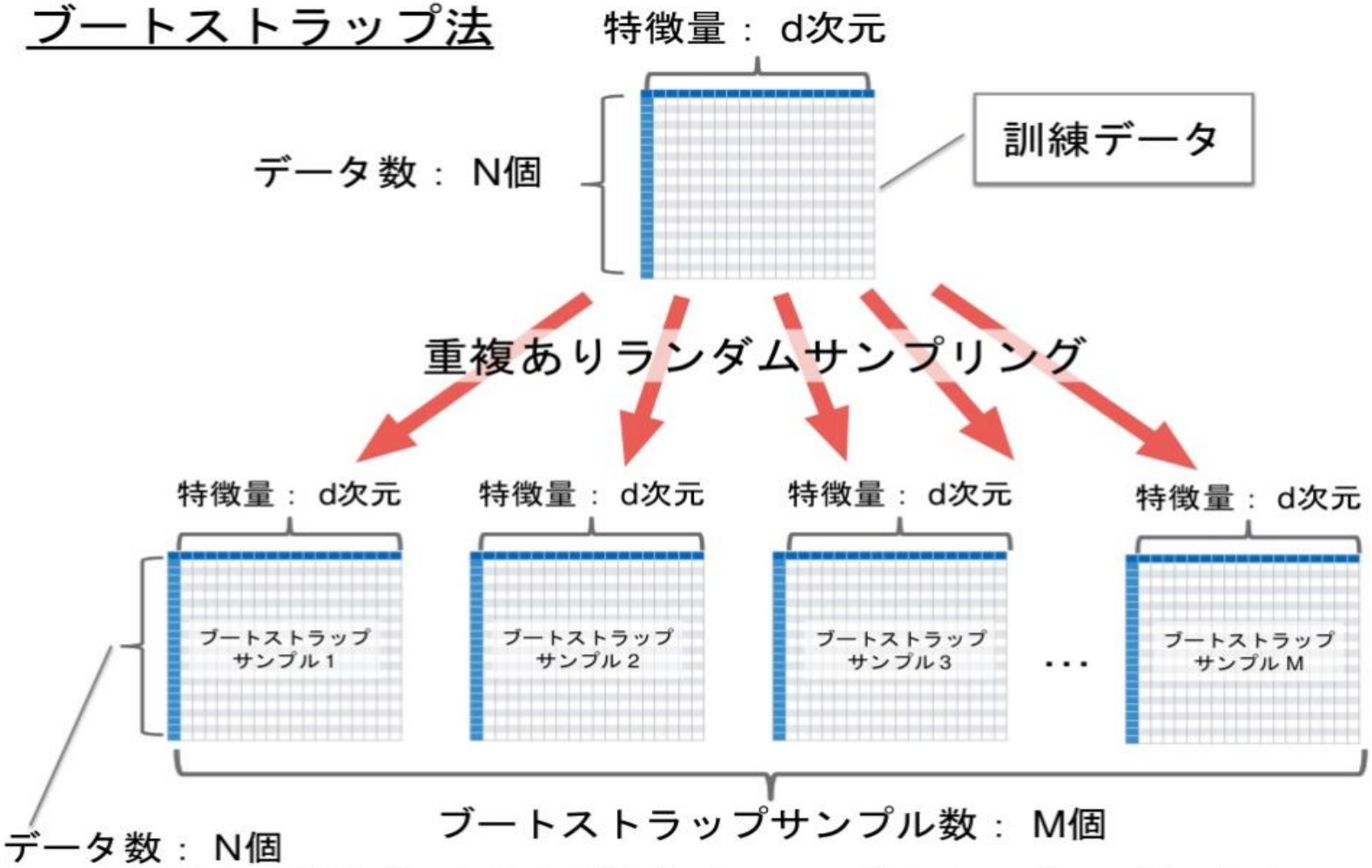
説明変数	目的変数
多群でも可	1群

- 説明変数…目的変数を説明する変数(物事の原因を説明)
- 目的変数…予測したい変数(物事の結果)

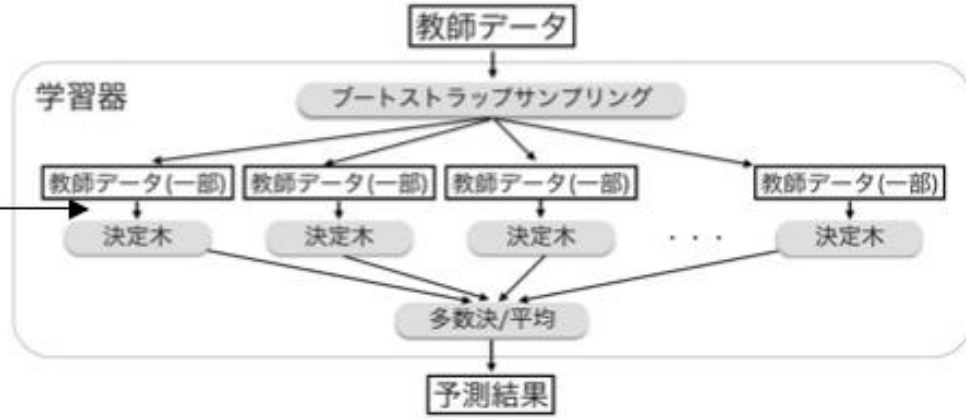
ランダムフォレストのイメージ

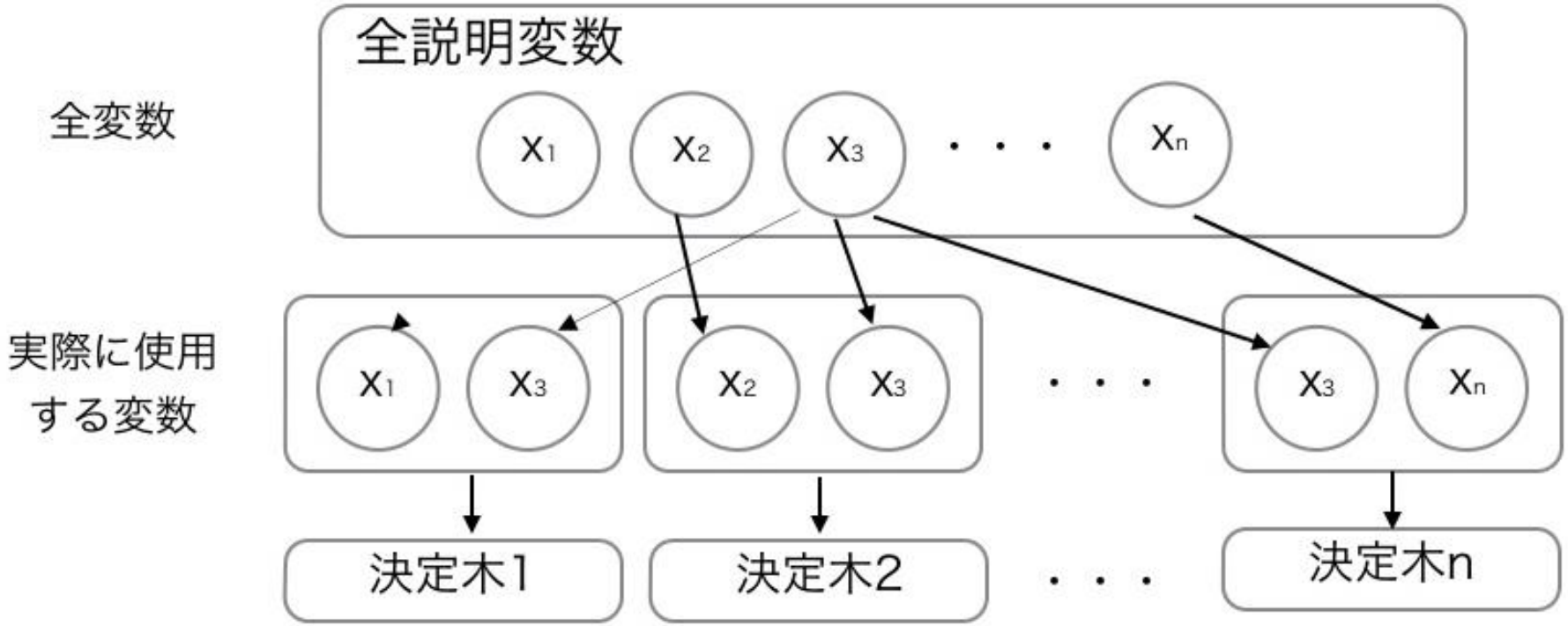


全教師データからランダムで抽出



学習の際、説明変数を
ランダムに選んで使う

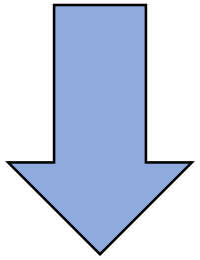




他の集団学習との違いは

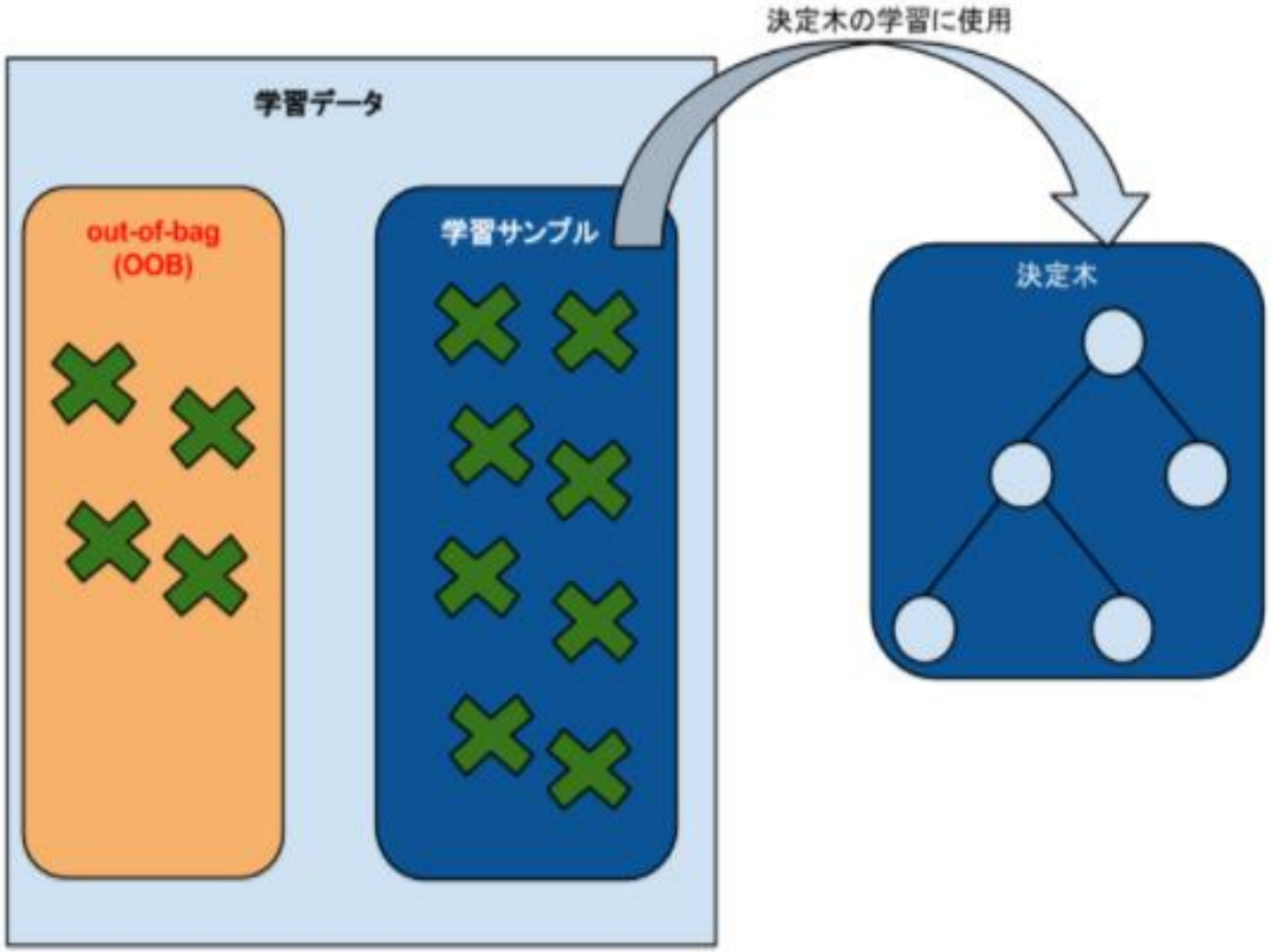
- 説明変数同士に相関があると、弱学習器間の相関が生まれる

弱学習器間に相関があるとバリアンスが下がらない



説明変数をランダムで選ぶことで、相関の低い決定木群を作成し
バリアンスを下げることができる

4.説明変数の重要度について



学習データをそのデータ数だけ
重複サンプリング

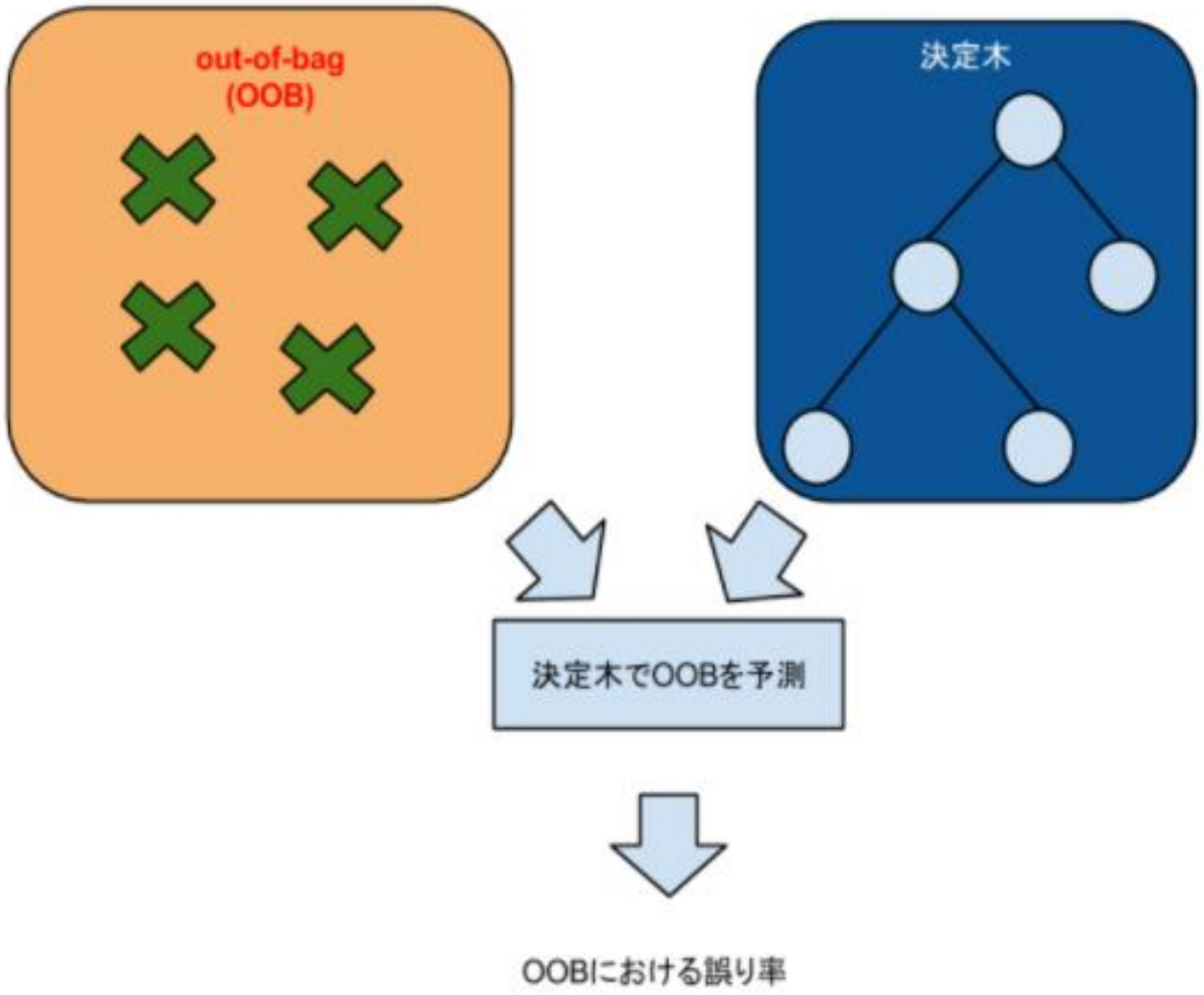
↓

決定木を学習させる

↓

このとき使われなかったデータ
を**out-of-bag（OOB）**という

OOBを説明変数の重要度に用いる



学習済みの決定木でその決定木のOOBを分類

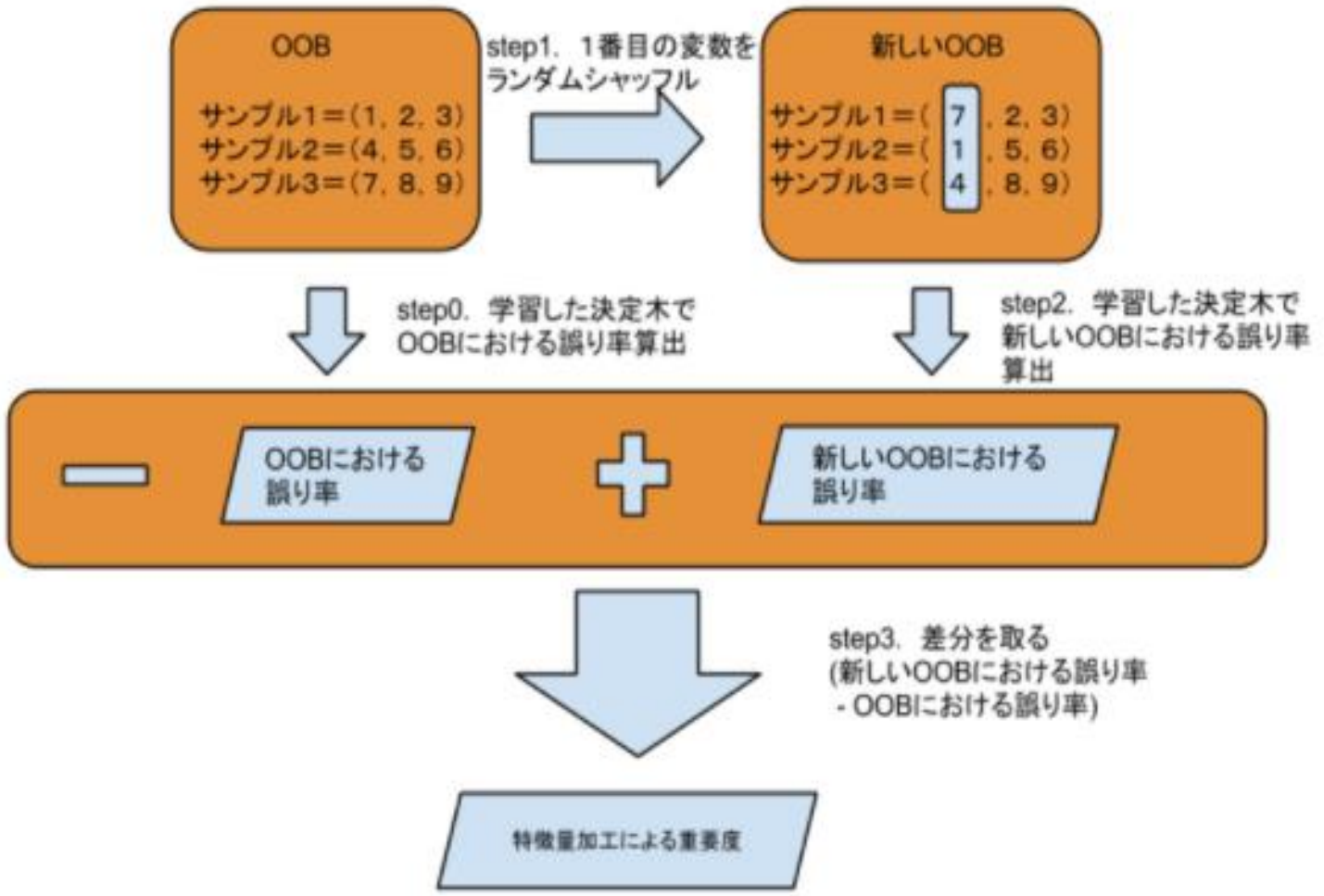
↓

OOBの各サンプルのうち間違って分類されたサンプルの割合

=

OOBの誤り率とする

誤り率を使って重要度を計算



OOBにおける誤り率を計算

↓

どの程度精度が下がったかを指標とする

↓

各決定木で行い、木あたりの平均を求める

=

この値が特徴量の重要度を表す

4.従来の手法と比較して

- **メリット**
  →ノイズに強い
  →過学習を防ぐことができる
  →説明変数の重要度が学習とともに計算できる

- **デメリット**
  →パラメータが多い(木の数や使用する説明変数の数)
  →学習データ/説明変数をランダム抽出するので、データと変数が少なすぎるとうまく学習できない。

まとめ

- ランダムフォレストとはどのような手法か知ることができた
- ランダムフォレストの仕組みを理解することで集団学習のメリットが分かった
- 従来の手法に比べ、説明変数の重要度が学習とともに計算できるため、結果の改善が期待できる