

半教師あり学習を用いた 深発月震の再分類の検討

中島康平¹，山本幸生²，山田竜平³，廣田雅春⁴，荒木徹也⁵，石川博⁶

1. 首都大学東京修士1年 2. 宇宙航空開発機構宇宙科学研究所
3. 会津大学 4. 岡山理科大学 5. 群馬大学 6. 首都大学東京

- ・ 研究背景・目的
- ・ 再分類手法
- ・ 実験結果

- ・ 研究背景・目的
- ・ 再分類手法
- ・ 実験結果

・ 月震データ

NASAのアポロミッションで月に地震計を設置
約7年半データを取得(1969 ~ 1977年)

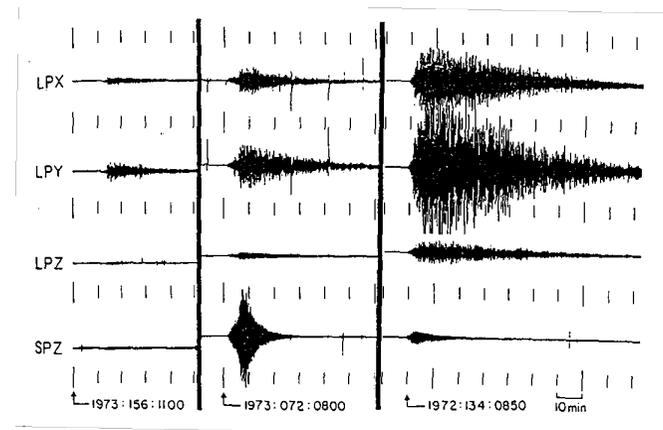
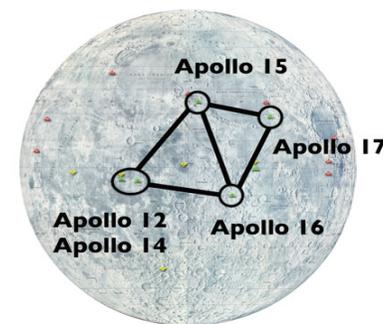
・ 月震の種類

深発月震, 浅発月震, 熱月震, 隕石衝突 の4種類
観測された月震の半数以上が深発月震

→ **本研究の対象も深発月震**

・ 深発月震の特徴

- 特定の震源から周期的に揺れが発生
- 同一の震源の月震は波形が類似
- ノイズが多く, 揺れが微弱



月震の波形

- ・ 月震を解析する理由

 - 月の内部構造の解明

 - 月震の発生メカニズムの理解

- ・ 月震解析に必要な情報

 - 月震の震源の特定

- ・ 月震の震源の特定

 - 3つの観測点で月震を観測することで震源を特定可能

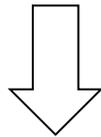
 - しかし、**ノイズ・振動の微弱さから、3点で観測されていない月震が多数存在**

- 2003年 月震の波形から震源を推定 (Nakamura et al., 2003)
 - 月震波形に対して相互相関関数を用いて震源を推定
 - 現在の月震分類の基準となっている
- 2014年 SOMを用いた月震のクラスタリング (Goto et al., 2014)
 - 特徴量にパワースペクトル密度を用いる
 - 従来の分類基準による分類結果の誤りを示唆
 - 未知の震源の存在を示唆

- 現在の問題点
 - 既存の分類結果が誤っている可能性がある
 - 未だ発見されていない未知の震源が存在する可能性がある

- 現在の問題点

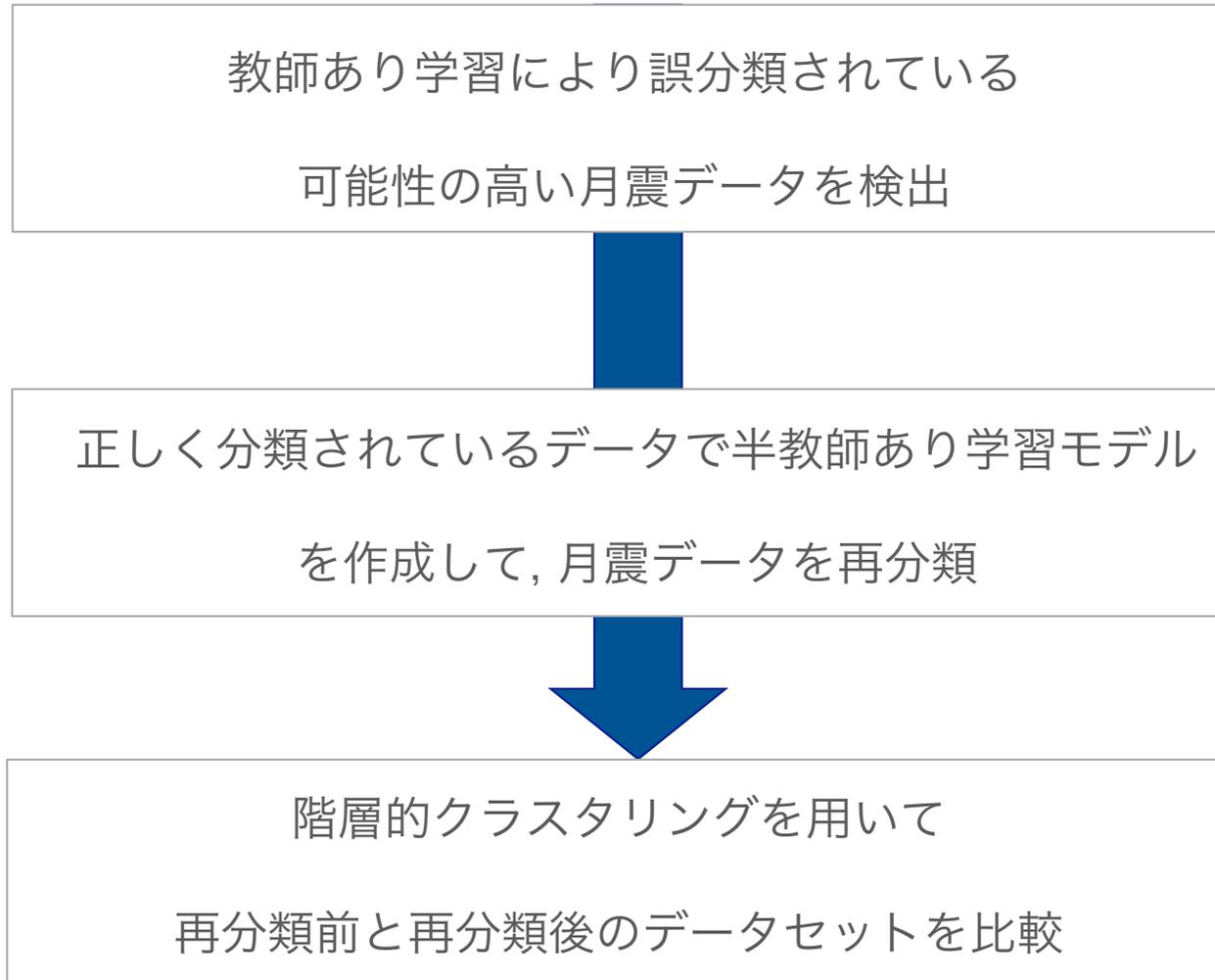
- 既存の分類結果が誤っている可能性がある
- 未だ発見されていない未知の震源が存在する可能性がある



- 半教師あり学習を用いた月震データの再分類

- 機械学習の基準に基づいた月震の震源ラベルを付与
- 未知の震源である可能性の高い月震を検出

- ・ 研究背景・目的
- ・ **再分類手法**
- ・ 実験結果



教師あり学習により誤分類されている

可能性の高い月震データを検出

正しく分類されているデータで半教師あり学習モデル

を作成して、月震データを再分類

階層的クラスタリングを用いて

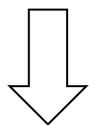
再分類前と再分類後のデータセットを比較

- 分類が正しいと考えられる月震データ

既存の分類結果と機械学習による推定が**一致**

- 分類が誤っていると考えられる月震データ

既存の分類結果と機械学習による推定が**不一致**

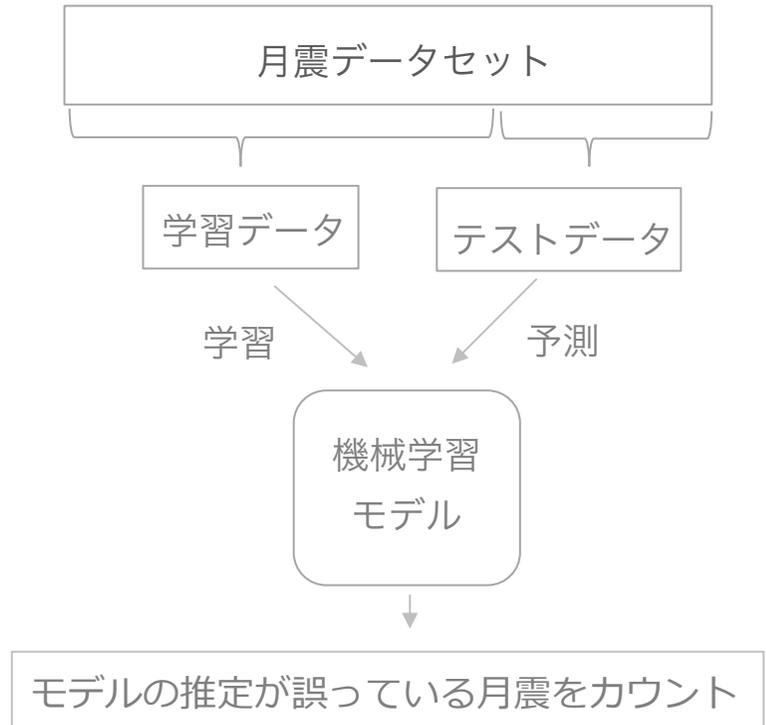


教師あり学習による分類を繰り返し行い、モデルの誤分類確率が高い月震データをラベルが誤っている可能性があると考え

誤分類の可能性のある月震の検出の流れ

1. データセットを学習データとテストデータに分割
2. 学習データを用いて、機械学習モデルの学習を行う
3. 学習後のモデルでテストデータの震源を予測
4. テストデータの中で、分類が誤っている月震を数える

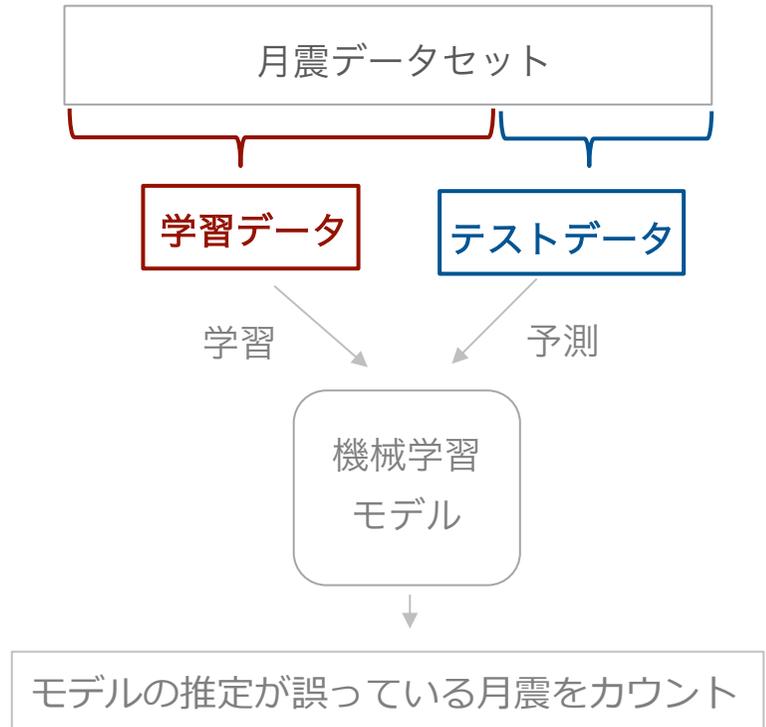
1～4の処理を繰り返し、全ての月震データに対する機械学習モデルの誤分類確率を計算



誤分類の可能性のある月震の検出の流れ

1. データセットを学習データとテストデータに分割
2. 学習データを用いて、機械学習モデルの学習を行う
3. 学習後のモデルでテストデータの震源を予測
4. テストデータの中で、分類が誤っている月震を数える

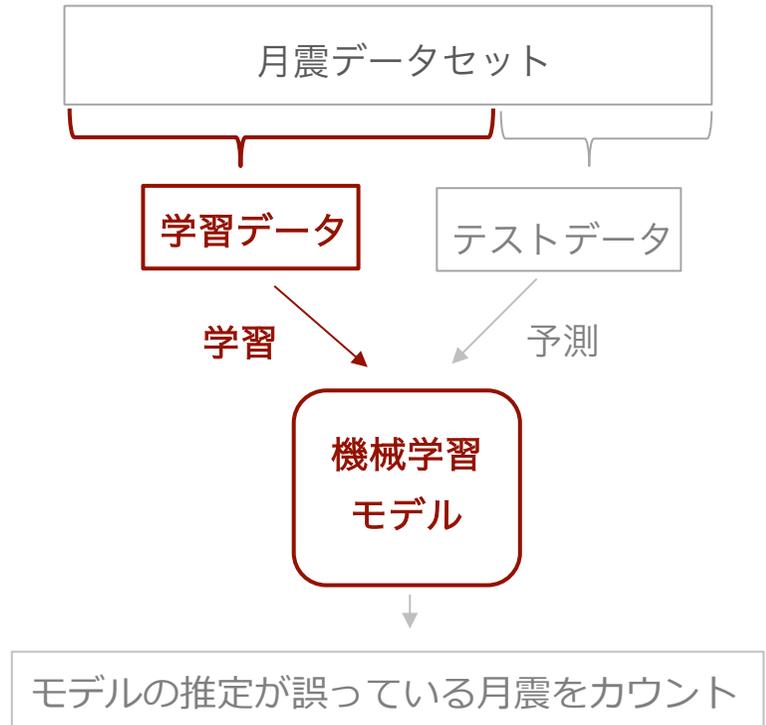
1～4の処理を繰り返し、全ての月震データに対する機械学習モデルの誤分類確率を計算



誤分類の可能性のある月震の検出の流れ

1. データセットを学習データとテストデータに分割
2. 学習データを用いて、機械学習モデルの学習を行う
3. 学習後のモデルでテストデータの震源を予測
4. テストデータの中で、分類が誤っている月震を数える

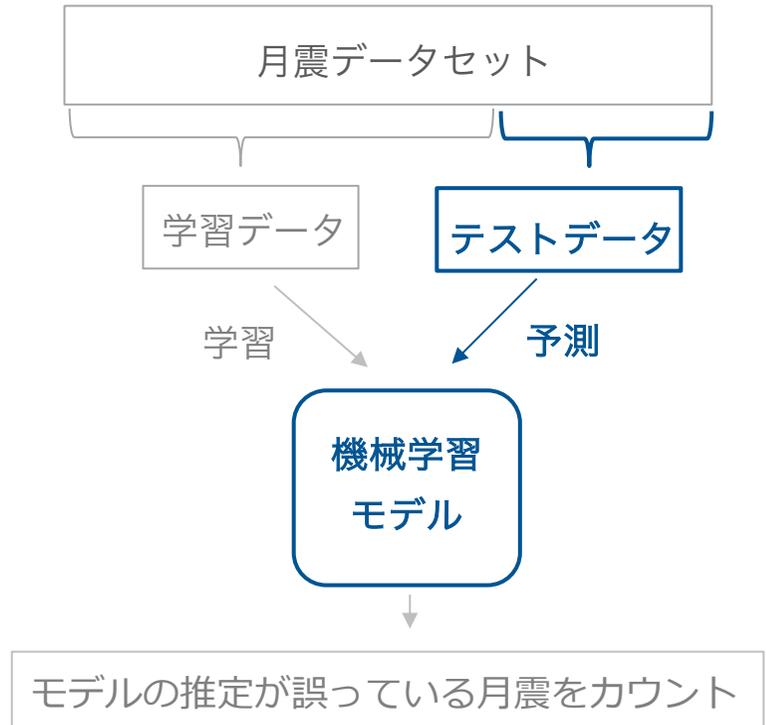
1～4の処理を繰り返し、全ての月震データに対する機械学習モデルの誤分類確率を計算



誤分類の可能性のある月震の検出の流れ

1. データセットを学習データとテストデータに分割
2. 学習データを用いて、機械学習モデルの学習を行う
3. 学習後のモデルでテストデータの震源を予測
4. テストデータの中で、分類が誤っている月震を数える

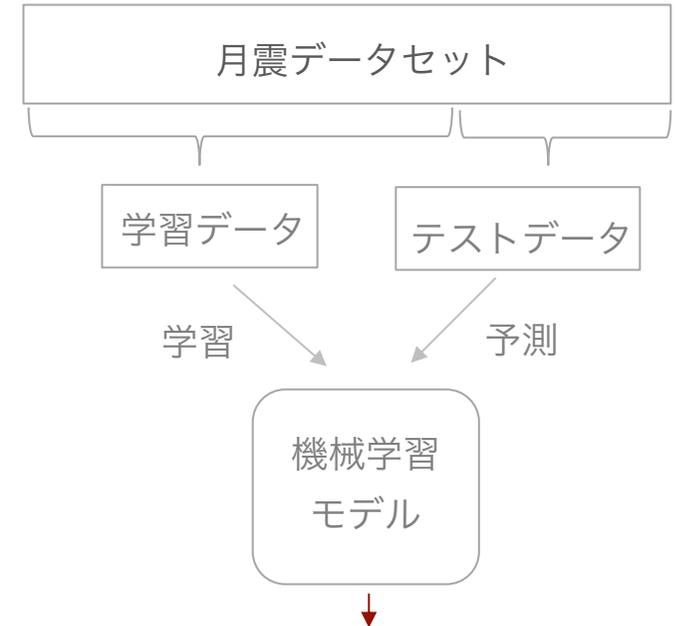
1～4の処理を繰り返し、全ての月震データに対する機械学習モデルの誤分類確率を計算



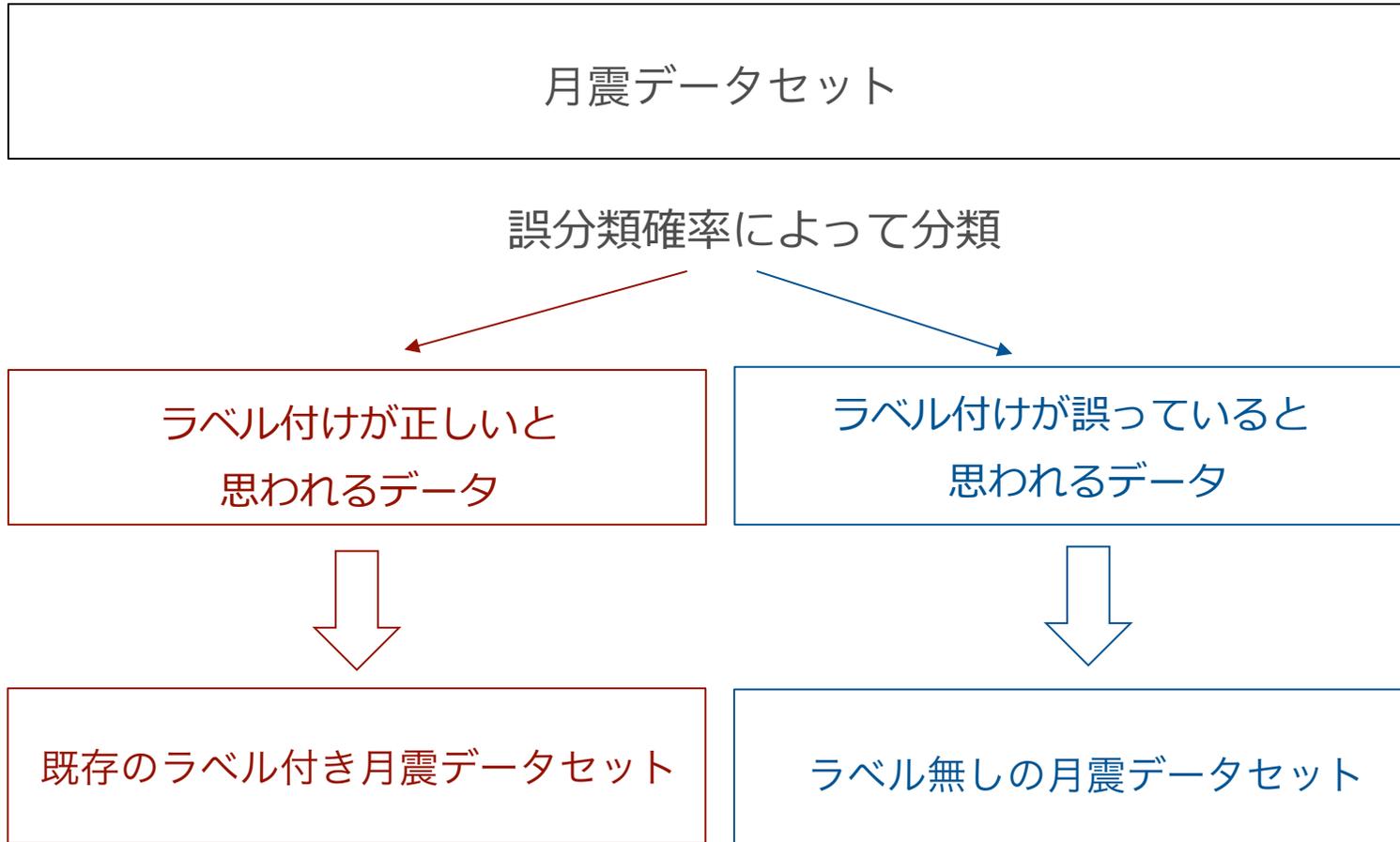
誤分類の可能性のある月震の検出の流れ

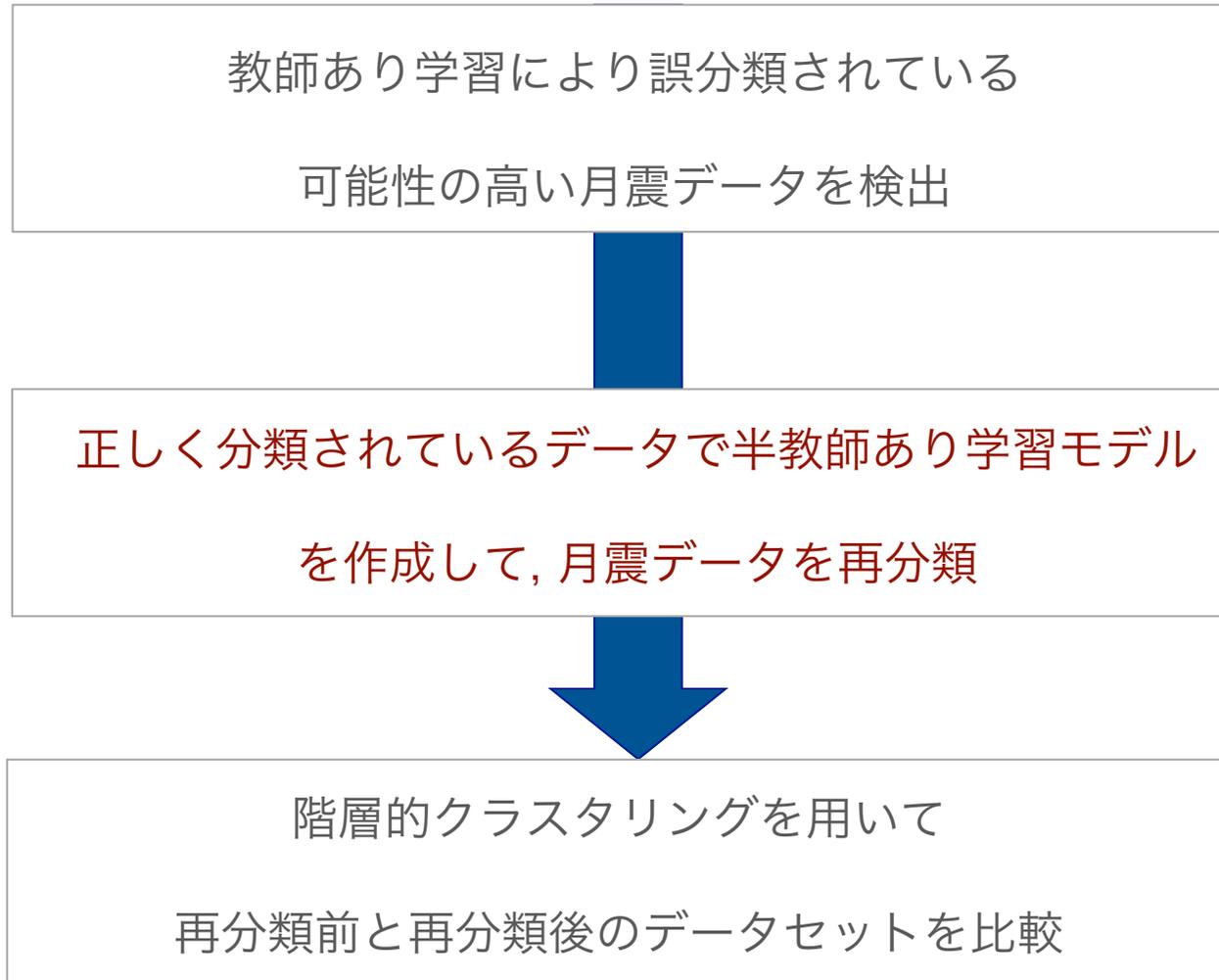
1. データセットを学習データとテストデータに分割
2. 学習データを用いて、機械学習モデルの学習を行う
3. 学習後のモデルでテストデータの震源を予測
4. テストデータの中で、分類が誤っている月震を数える

1～4の処理を繰り返し、全ての月震データに対する機械学習モデルの誤分類確率を計算



モデルの推定が誤っている月震をカウント

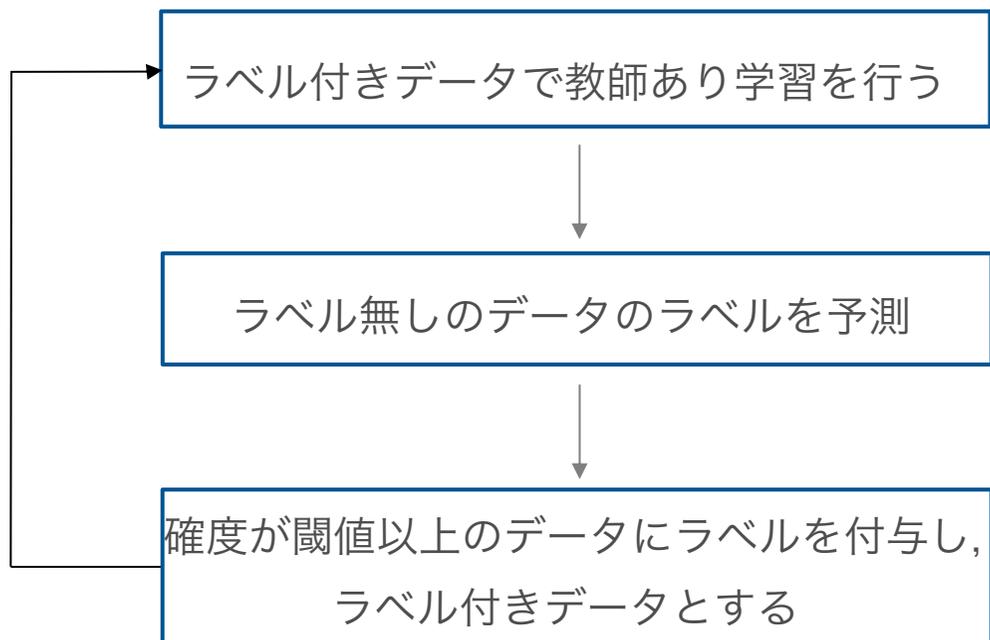




- 半教師あり学習

一部のラベル付きデータから、ラベル無しデータのラベルを予測するモデル

- Self Training



- 分類器

ロジスティック回帰

- 分類に用いられる機械学習モデル
- 入力データがどのクラスに分類されるかを**確率で出力** → 確率を確度とする

- Self Trainingを用いた月震の再分類

1. 特定の震源への確率(確度)が95%以上の月震データにラベルを付与
2. ラベルを付与できる月震データが一つも存在しない場合, 確度の閾値を5%下げる
3. 確度の閾値が50%以下になった場合, ラベル付けを終了
4. ラベルを付与できた場合, 確度の閾値は95%に戻す

- 分類器

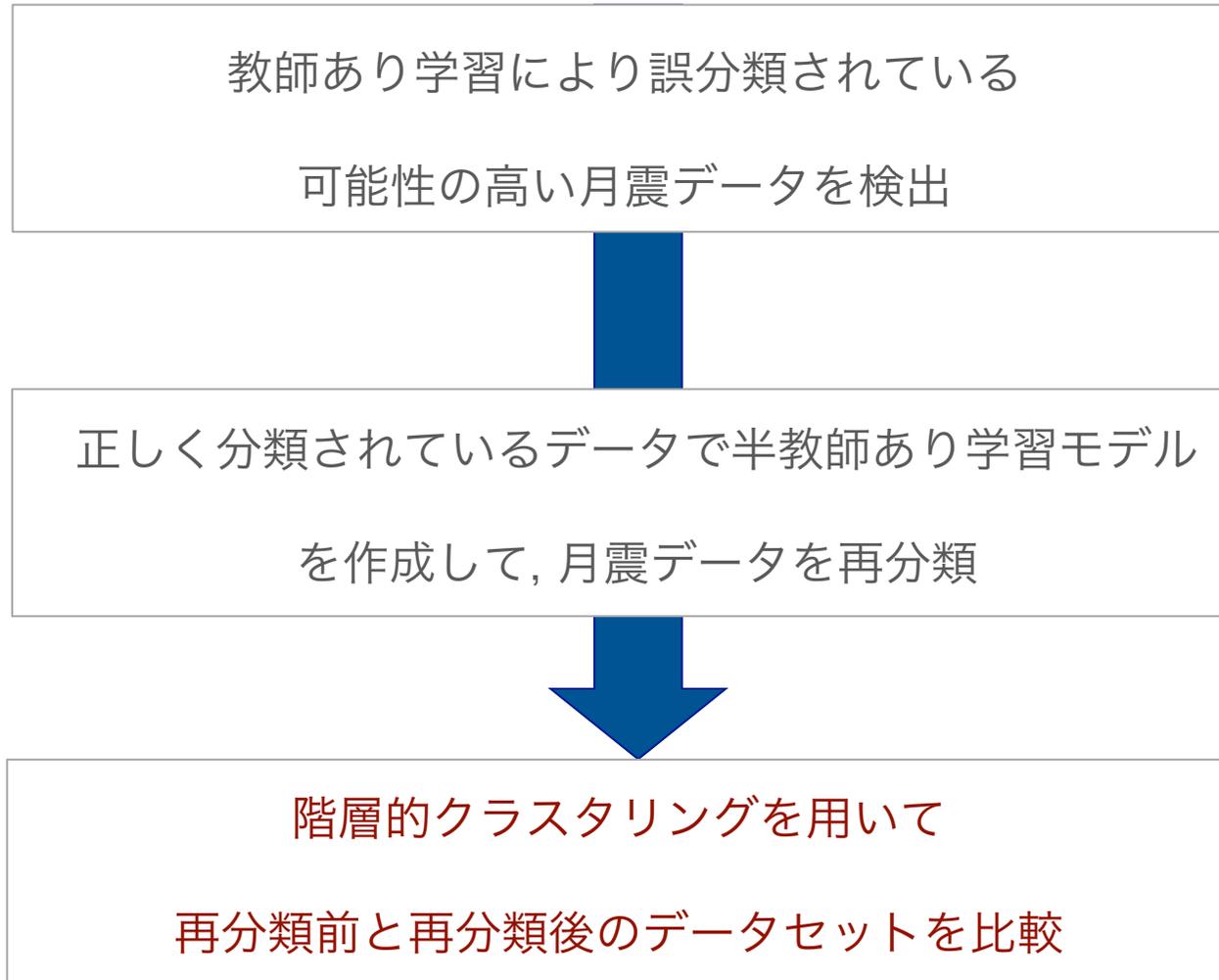
ロジスティック回帰

- 分類に用いられる機械学習モデル
- 入力データがどのクラスに分類されるかを**確率で出力** → 確率を確度とする

- Self Trainingを用いた月震の再分類

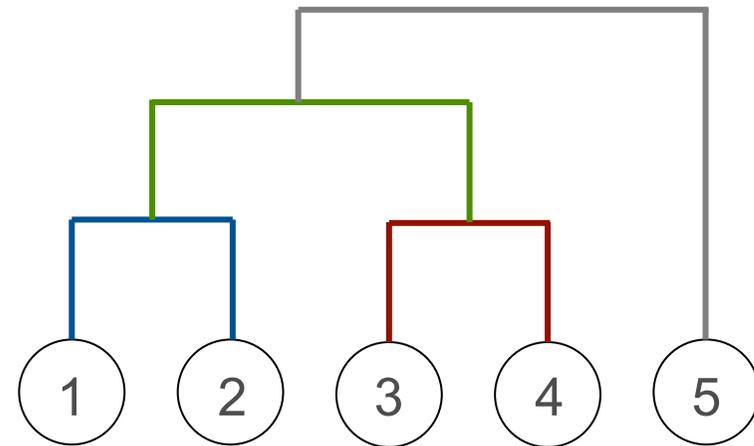
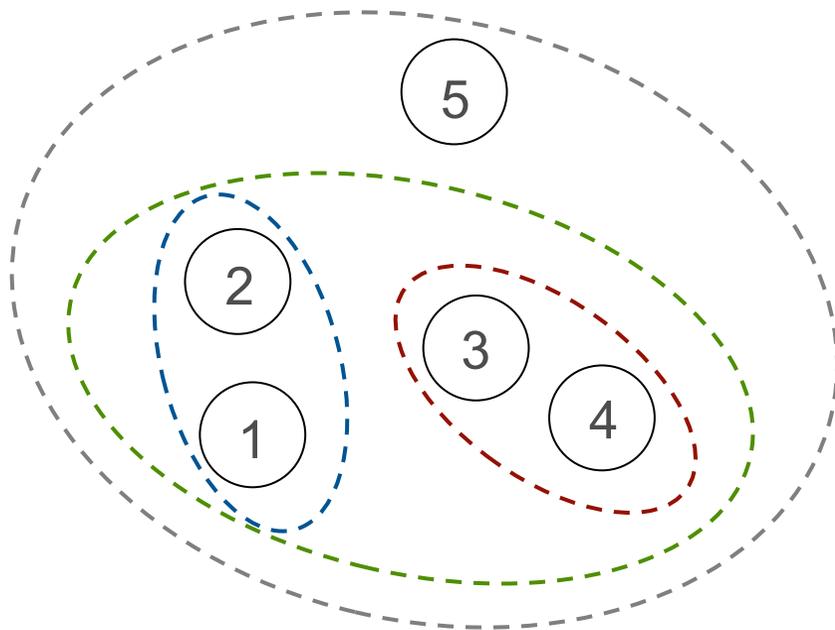
1. 特定の震源への確率(確度)が95%以上の月震データにラベルを付与
2. ラベルを付与できる月震データが一つも存在しない場合, 確度の閾値を5%下げる
3. **確度の閾値が50%以下になった場合, ラベル付けを終了**
4. ラベルを付与できた場合, 確度の閾値は95%に戻す

ラベルが付与されていない月震データは既存の震源には、
当てはまらない可能性のある月震データとする



・ 階層的クラスタリング

1. 全てのデータそれぞれを一つのクラスタとする
2. 距離関数に基づき、もっとも近い2つのクラスタを併合
3. クラスタが一つになるまでクラスタを併合



- 研究背景・目的
- 再分類手法
- **実験結果**

- **実験に用いたデータ**

データ数が50件以上の震源の月震データ
(AP12で観測された月震のみ)

- **前処理**

1. オフセット処理
2. トレンド引き
3. バンドパスフィルタ
(0.3 – 1.5 Hz)
4. デスパイキング処理

ノイズ除去

- **特徴量**

Z軸方向

振動開始から15分の周波数スペクトル

震源	データ数
A1	261
A6	85
A8	93
A9	94
A10	108
A14	87
A18	106
A20	106
A23	54
合計	994

震源ごとのデータ数

- データ分割割合

学習データ : テストデータ = 7 : 3

- 教師あり学習を行なった回数

ロジスティック回帰で1000回

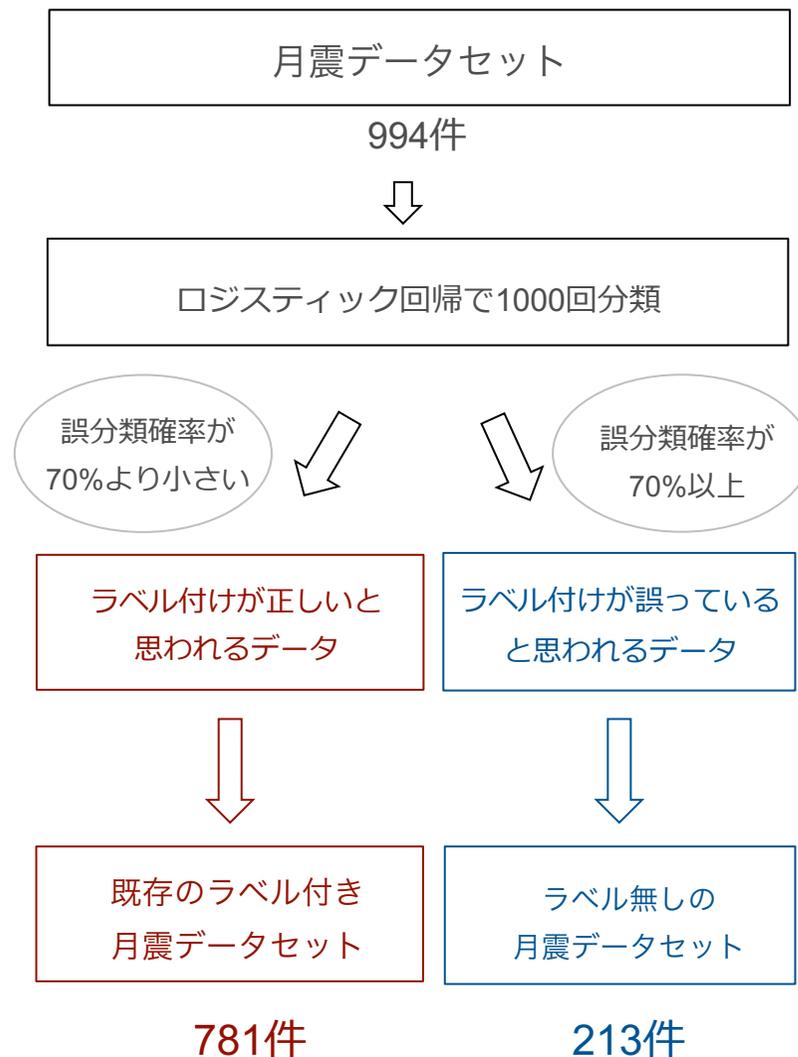
- ラベル無しとする月震

誤分類確率70%以上の月震

- 分類結果

既存のラベル付き月震データセット : 781件

ラベル無しの月震データセット : 213件



- 機械学習の分類器

- ロジスティック回帰

- Self training

1. 特定のクラスへの確度が95%以上のデータにラベルを付与
2. ラベルを付与できるデータが一つも存在しない場合, 確度の閾値を5%下げる
3. 確度の閾値が50%以下になった場合, ラベル付けを終了
4. ラベルを付与できた場合, 確度の閾値は95%に戻す

震源	データ数
A1	261
A6	85
A8	93
A9	94
A10	108
A14	87
A18	106
A20	106
A23	54
合計	994

初期の月震データセット

震源	データ数
A1	265
A6	65
A8	128
A9	84
A10	90
A14	73
A18	84
A20	73
A23	46
合計	909

再分類後の月震データセット

震源	データ数
A1	14
A6	22
A8	14
A9	1
A10	6
A14	7
A18	8
A20	10
A23	4
合計	85

どの震源にも当てはまらない
と推定されて月震データ数

震源	データ数
A1	261
A6	85
A8	93
A9	94
A10	108
A14	87
A18	106
A20	106
A23	54
合計	994

初期の月震データセット

震源	データ数
A1	265
A6	65
A8	128
A9	84
A10	90
A14	73
A18	84
A20	73
A23	46
合計	909

再分類後の月震データセット

震源	データ数
A1	14
A6	22
A8	14
A9	1
A10	6
A14	7
A18	8
A20	10
A23	4
合計	85

どの震源にも当てはまらない
と推定されて月震データ数

- クラスタ結合法

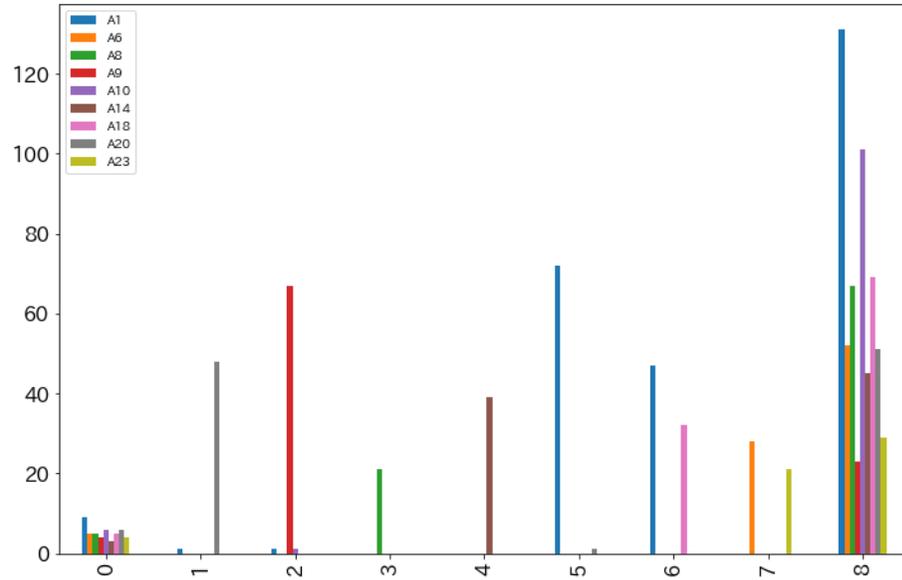
Ward法

- 距離の定義

標準ユークリッド距離

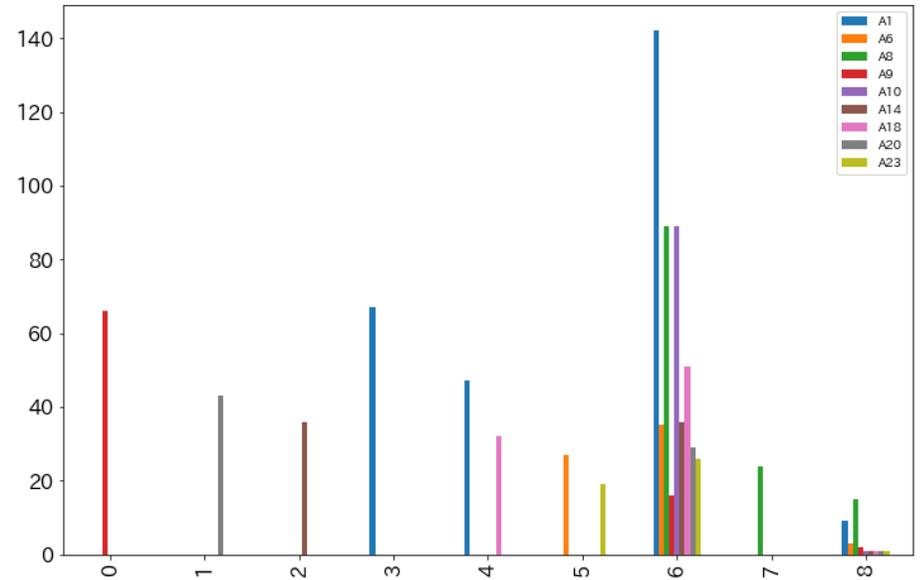
- クラスタ数

9 (震源数と同じ)



再分類前のデータセット

コーフェン相関係数 : 0.727



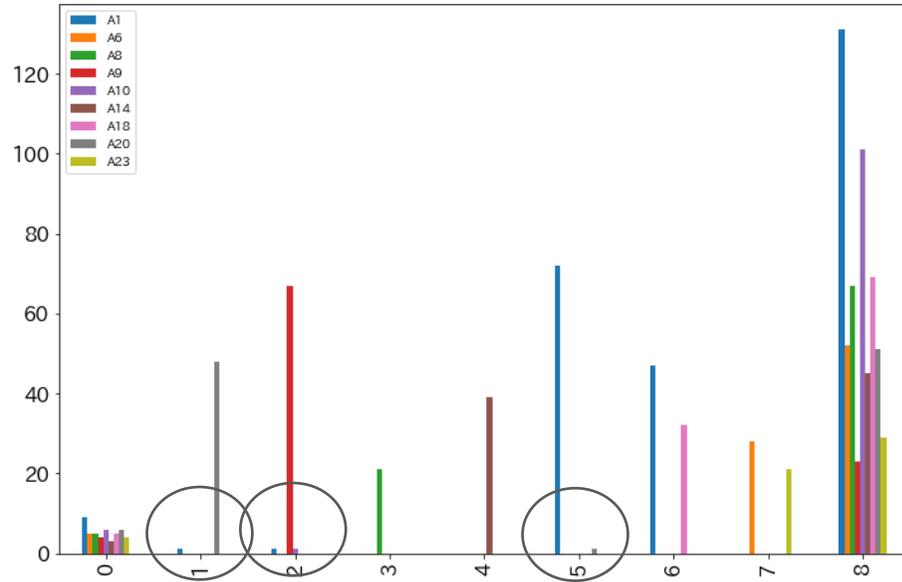
再分類後のデータセット

コーフェン相関係数 : **0.855**

- コーフェン相関係数

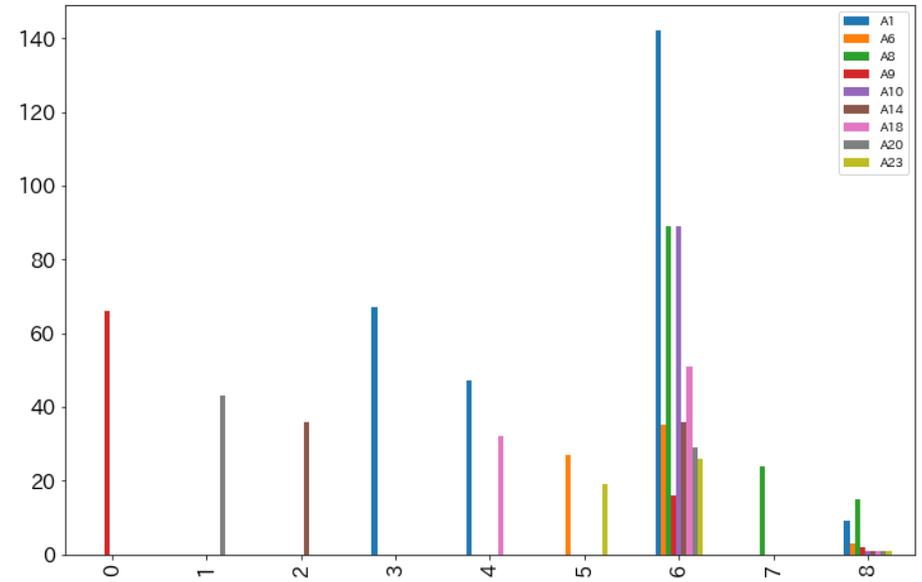
-1 ~ 1の値で算出され、1に近いほどクラスターの妥当性が高い

既存の震源に当てはまらないデータを排除したため、
データセットのまとまりが向上



再分類前のデータセット

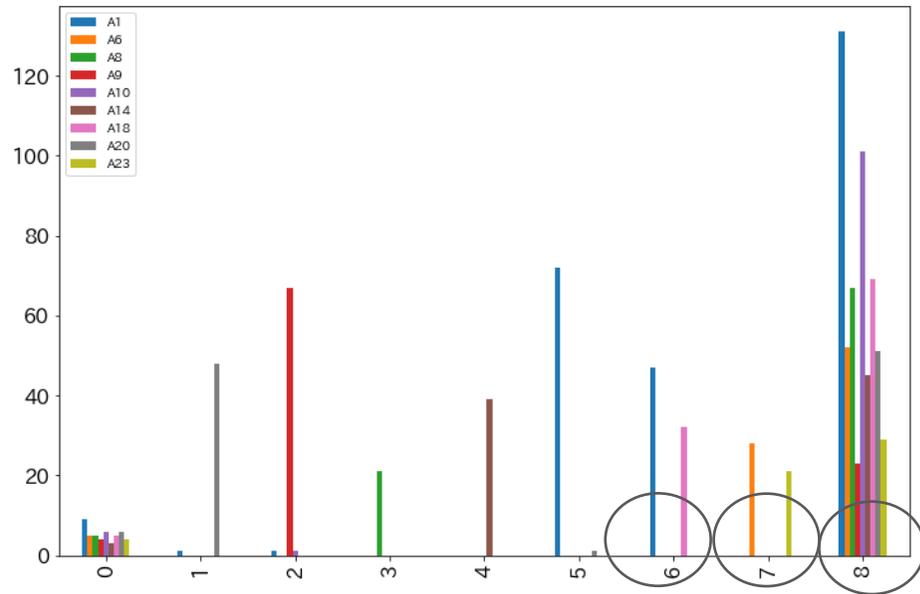
コーフェン相関係数 : 0.727



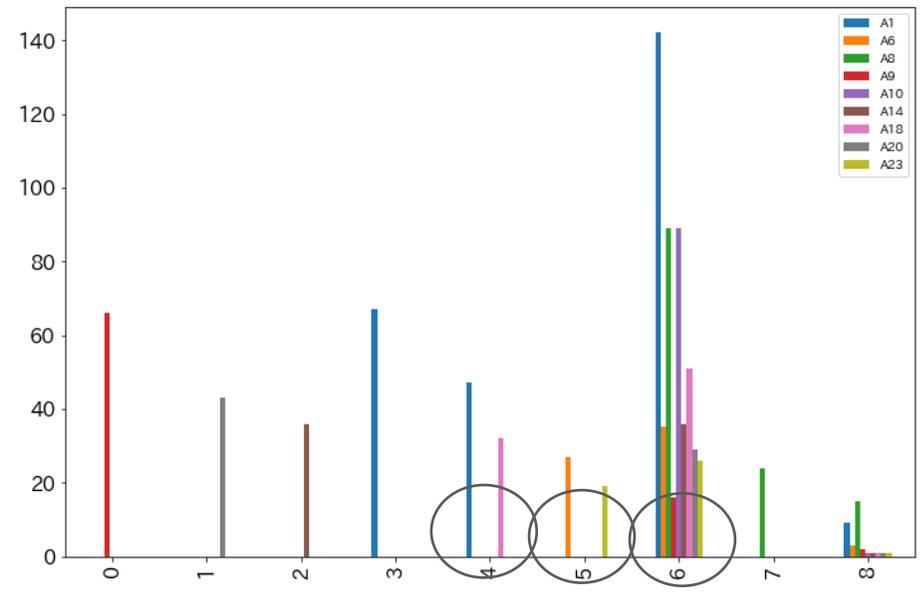
再分類後のデータセット

コーフェン相関係数 : 0.855

再分類により, 単一の震源からの月震データのみ
で構成されるクラスタが増加



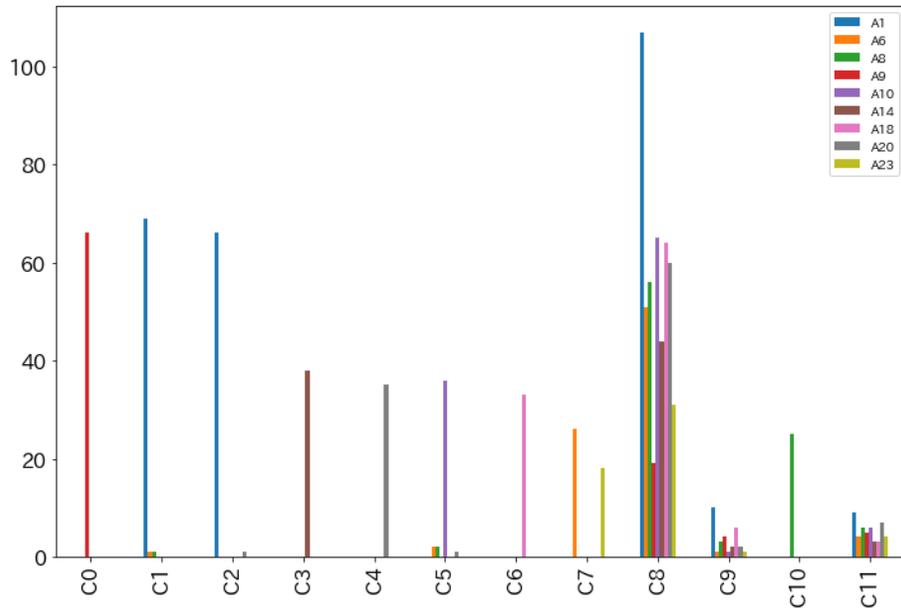
再分類前のデータセット



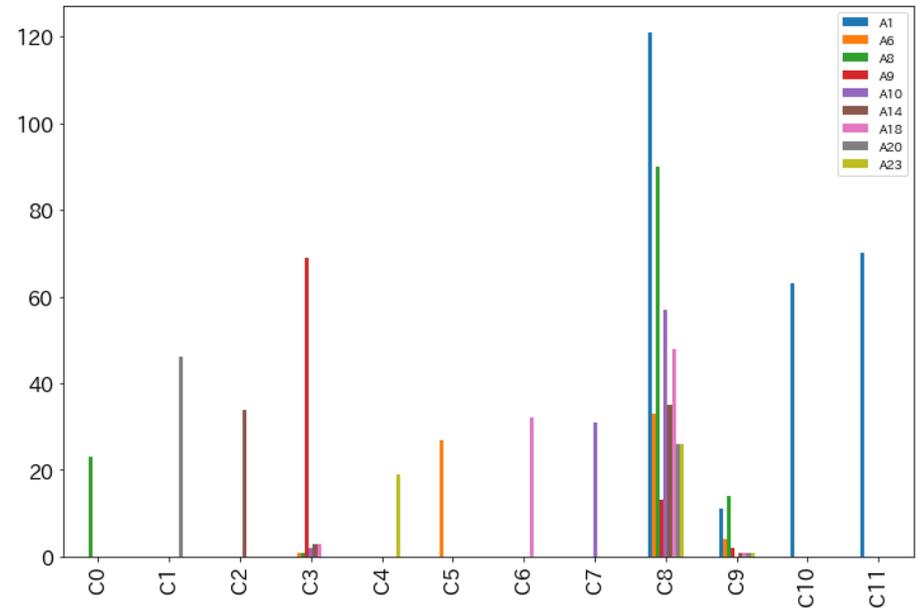
再分類後のデータセット

クラスタ数9の場合, 分離しきれていないクラスタが存在する

→ クラスタ数を12として再度クラスタリングを行う



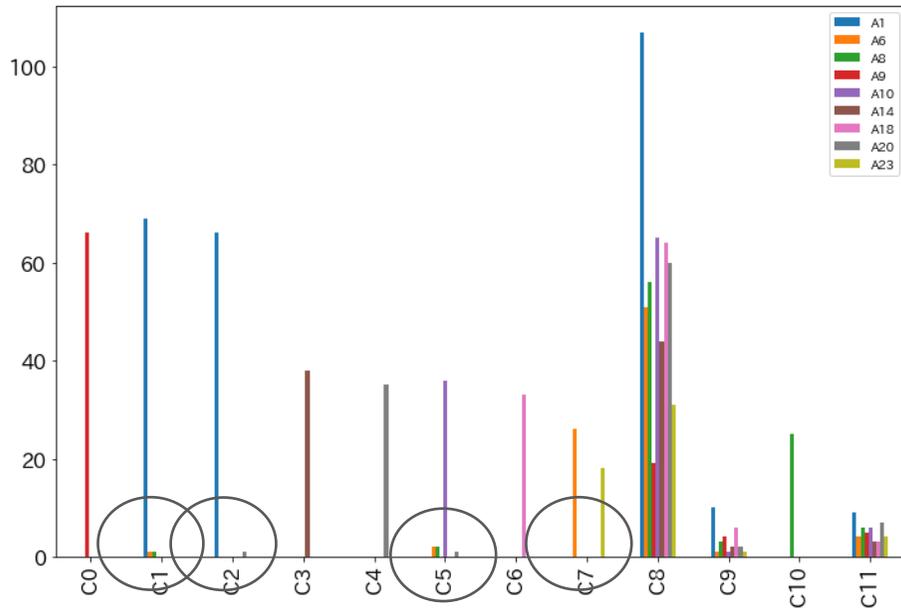
再分類前のデータセット



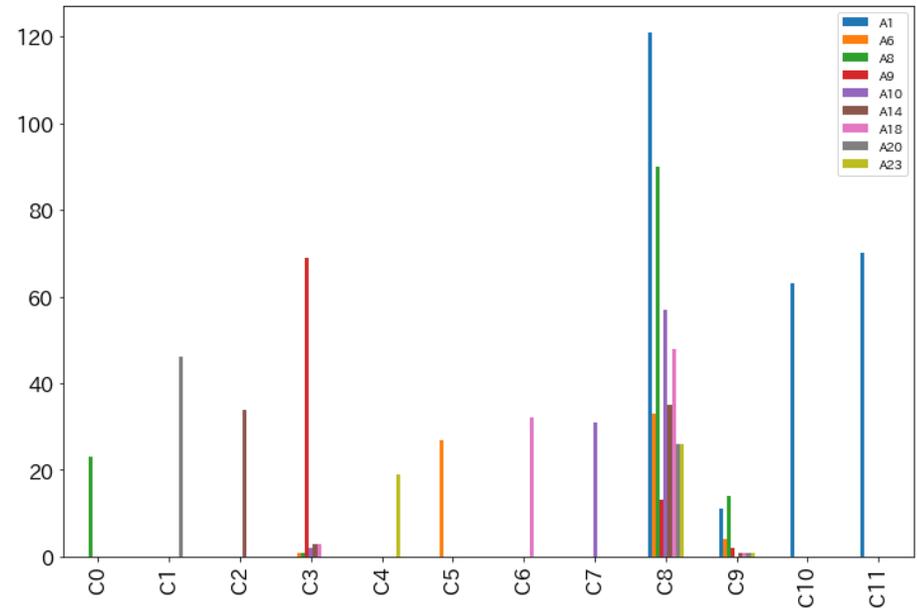
再分類後のデータセット

クラスタ数9の場合, 分離しきれていないクラスタが存在する

→ クラスタ数を12として結果を表示



再分類前のデータセット

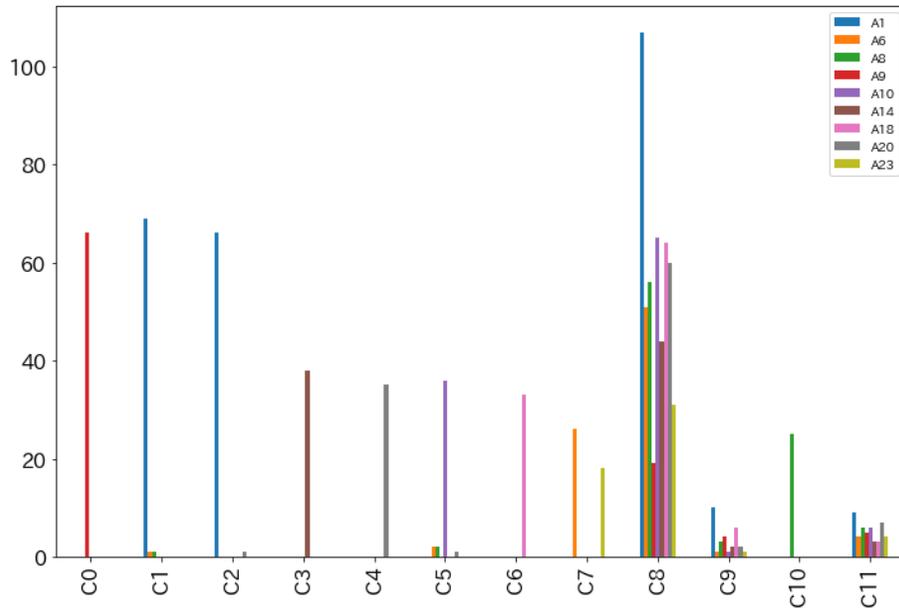


再分類後のデータセット

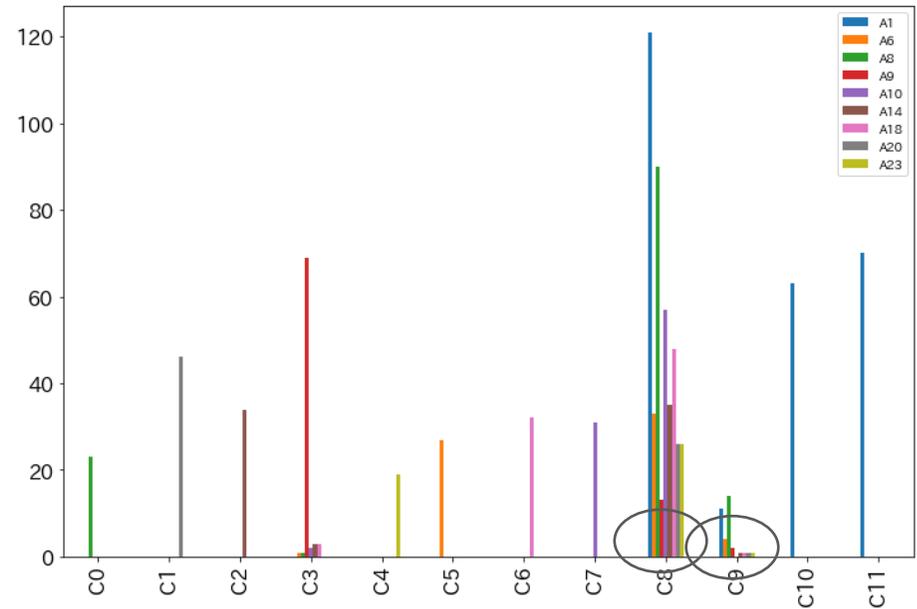
クラスタ数9の場合, 分離しきれていないクラスタが存在する

→ クラスタ数を12として結果を表示

再分類により, 単一の震源からの月震データのみ
で構成されるクラスタが増加



再分類前のデータセット



再分類後のデータセット

クラスター数9の場合, 分離しきれていないクラスターが存在する

→ クラスター数を12として結果を表示

再分類を行い, クラスター数を増やしても複数の震源から発生した月震データで構成されるクラスターが存在

- 再分類後は、データセットのまとまりが向上

データセットから排除した月震データの中に未知の震源から発生した月震が含まれている可能性がある

- 再分類後は、単一の震源から構成されるクラスタが増加

クラスタリング結果は、再分類後の方が望ましい結果となった

- 震源固有の特徴が弱い月震データが存在

再分類・データの排除を行っても、うまくクラスタリングできない月震データが存在する

- ・ どの震源にも当てはまらないとした月震データの分析
- ・ クラスタリング手法の再検討
- ・ 専門家による再分類後のデータセットの妥当性の診断

- **行ったこと**
 - 半教師あり学習を用いた深発月震の再分類
- **再分類の結果**
 - 単一の震源からの月震データのみで構成されるクラスタ数が増加
 - 再分類後のデータセットでも, 分類しきれない月震データが存在
- **今後の課題**
 - どの震源にも当てはまらないとした月震データの分析
 - クラスタリング手法の再検討
 - 専門家による再分類後のデータセットの妥当性の診断