

機械学習によるX線天体の分類手法

愛媛大学

高エネルギー天文学研究室

B4 菊池 誠

指導教員 寺島 雄一

1 目的

- 目的

機械学習によるX線天体の分類手法の開発



機械が自動的に学習し予測や分類を行う

X線天文衛星XMM-newtonによる100万を超えるX線源の観測

→ 大多数が未分類

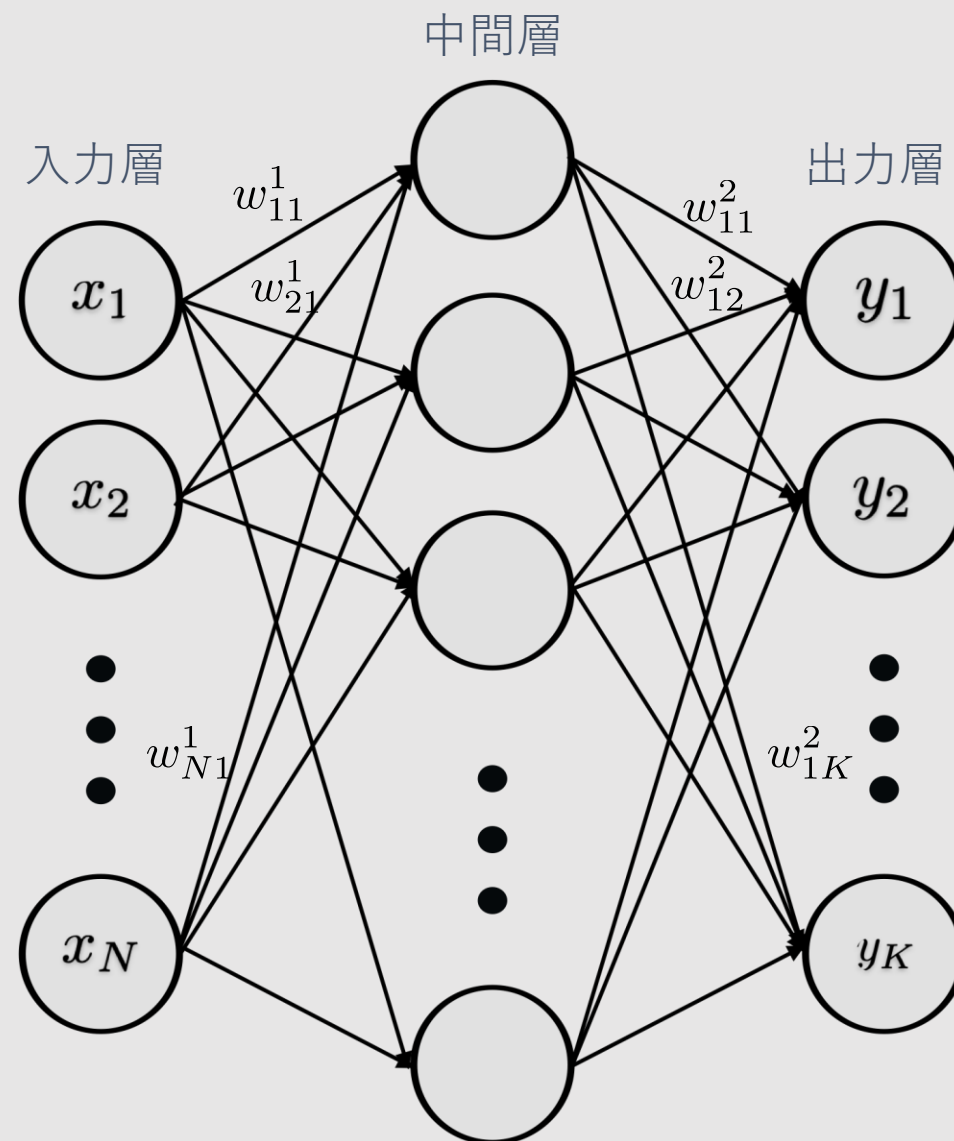
機械学習にやらしてもらおう!!

宇宙における高エネルギー現象の多くはX線で観測

2 手法

- 用いた手法
ニューラルネットワーク
教師あり学習
- ニューラルネットワークとは
脳の神経細胞(ニューロン)の
つながりを単純化し数理的に
表したモデル
入力層、中間層(隠れ層)、出力層の3種

重み付き線形和($\sum x_i w_i^l$)を入力
出力層では確率分布を出力



ニューラルネットワーク(三層)の模式図
今回用いたのは中間層2層

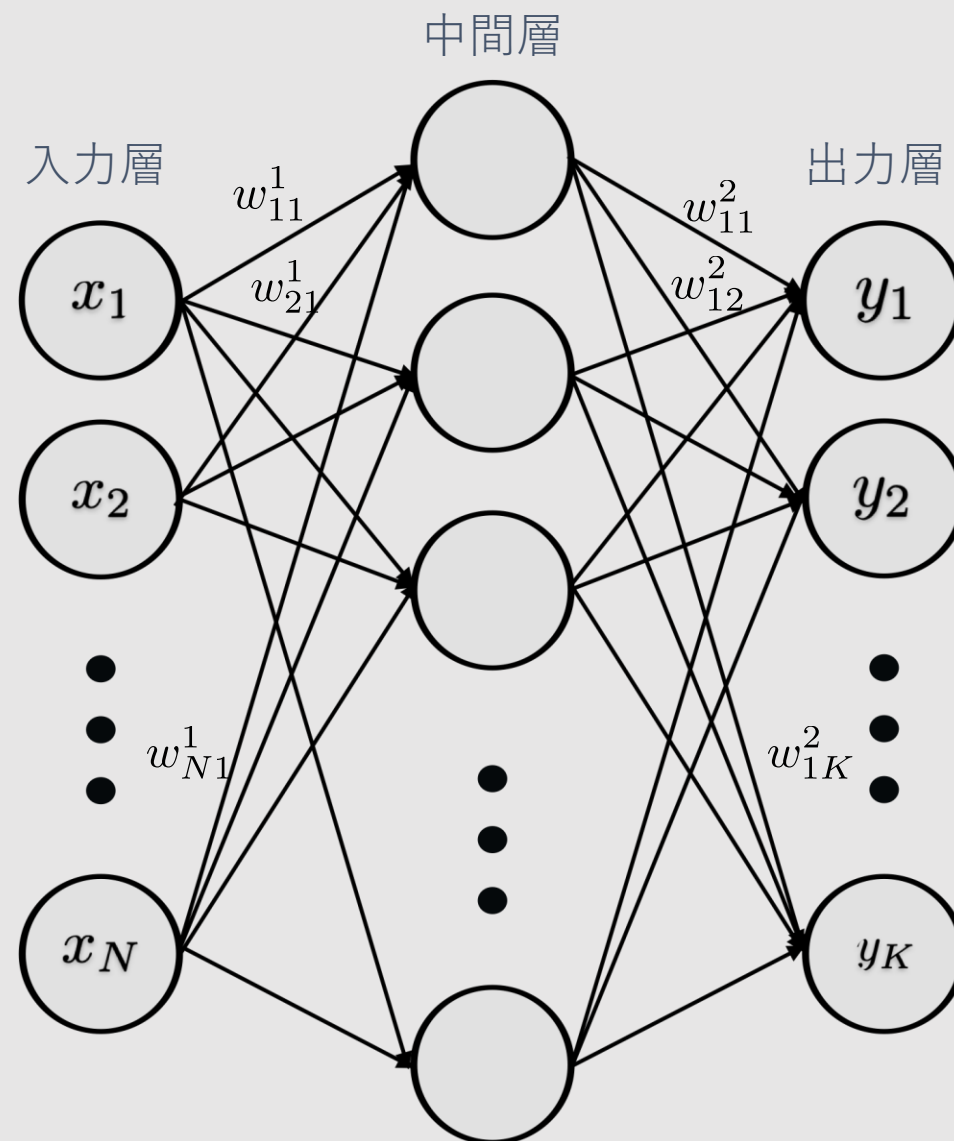
2 手法

- 教師あり学習
入力データ(特徴量)と
答え(クラスラベル)の関係を学習
→教師データ

予想結果と正解の差を最小に
するように重みを最適化

↓
学習

- 多層パーセプトロンを使用



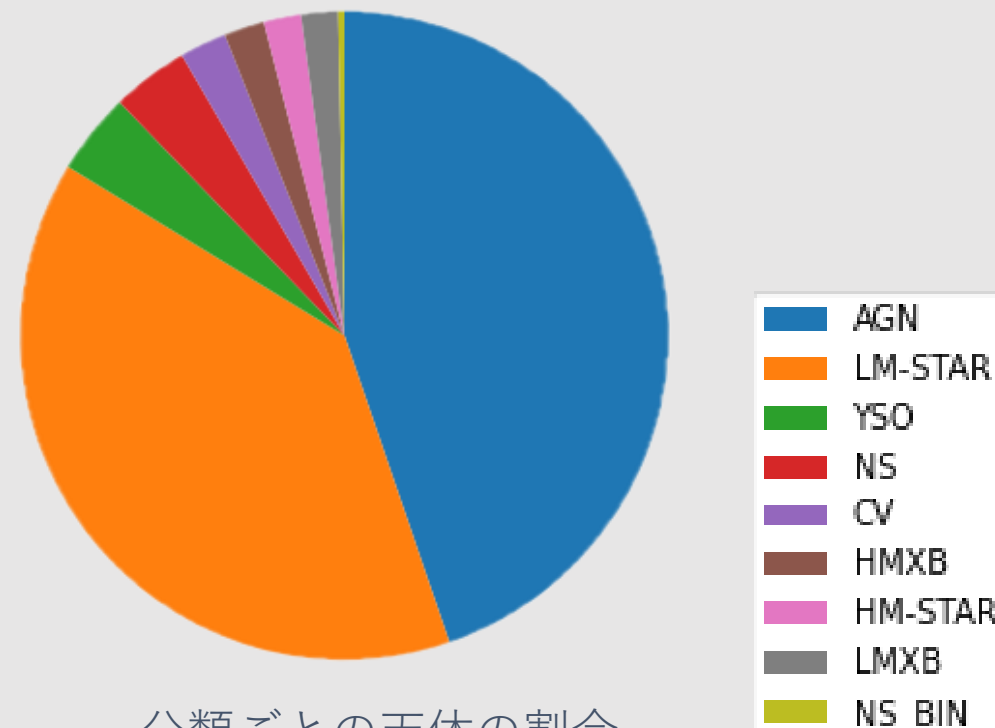
ニューラルネットワーク(三層)の模式図
今回用いたのは中間層2層

2 手法 教師データ

X線天文衛星XMM-Newtonによって観測されたX線源

『**Multiwavelength Catalog of 10,000 4XMM-DR13 Sources with Known Classifications**』 (Lin et al.2024)

AGN	:活動銀河核(7153)
CV	:激変星(373)
HM-STAR	:大質量星(299)
LM-STAR	:小質量星(6231)
HMXB	:大質量X線連星(322)
LMXB	:小質量X線連星(292)
YSO	:原始星(650)
NS	:パルサー、中性子星(614)
NS_BIN	:特殊なパルサーなど(50)



分類ごとの天体の割合

2 手法 特徴量

Count Rate (counts/s)

0.2 - 12keVにおけるカウントレート

HR 1 - 4: Hardness ratio

バンド1 - 5の硬度比

バンド1 : 0.2 - 0.5keV バンド2 : 0.5 - 1.0keV

バンド3 : 1.0 - 2.0keV バンド4 : 2.0 - 4.5keV

バンド5 : 4.5 - 12.0keV

Chi2prob (欠損の有無)

フラックスが一定である確率

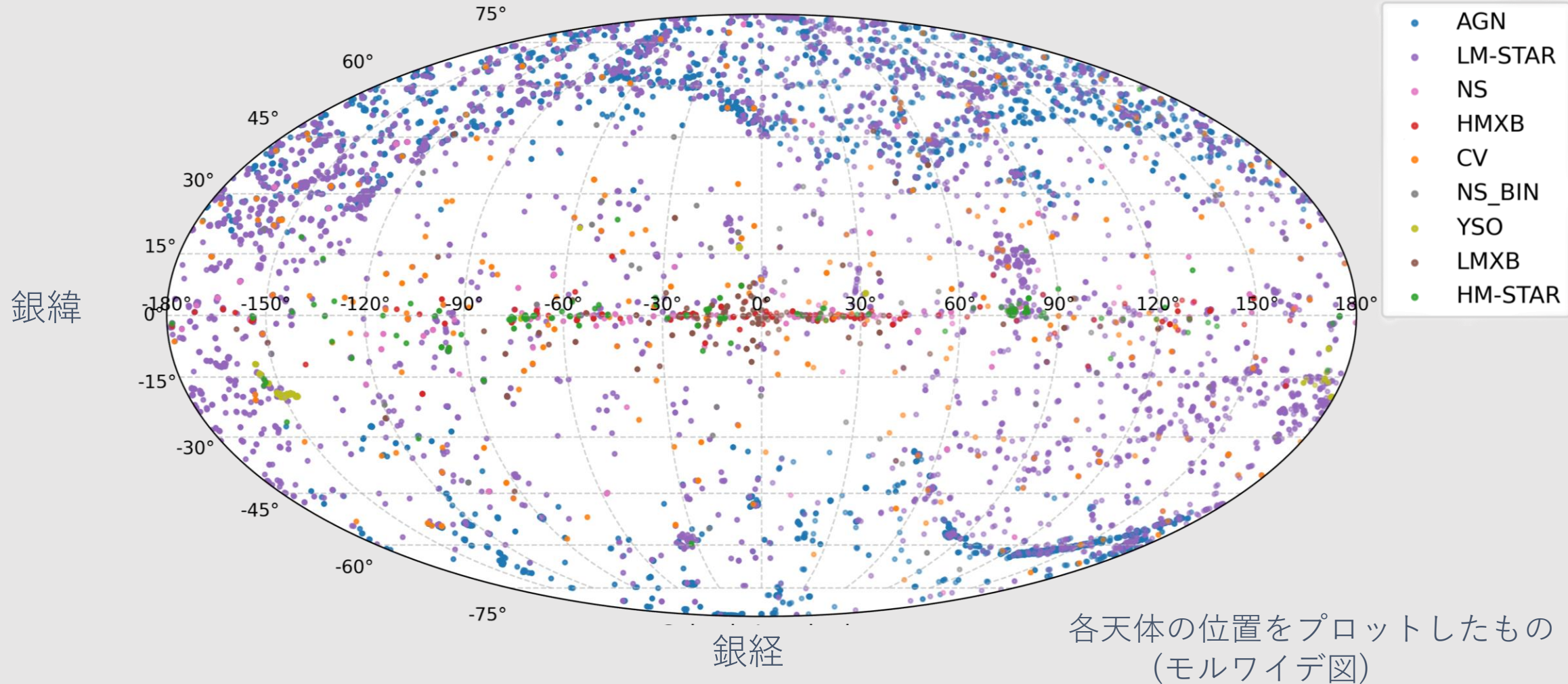
**Fvar : fractional RMS variability
amplitude (欠損の有無)**

変動のRMS

Extent parameter

見かけの広がり

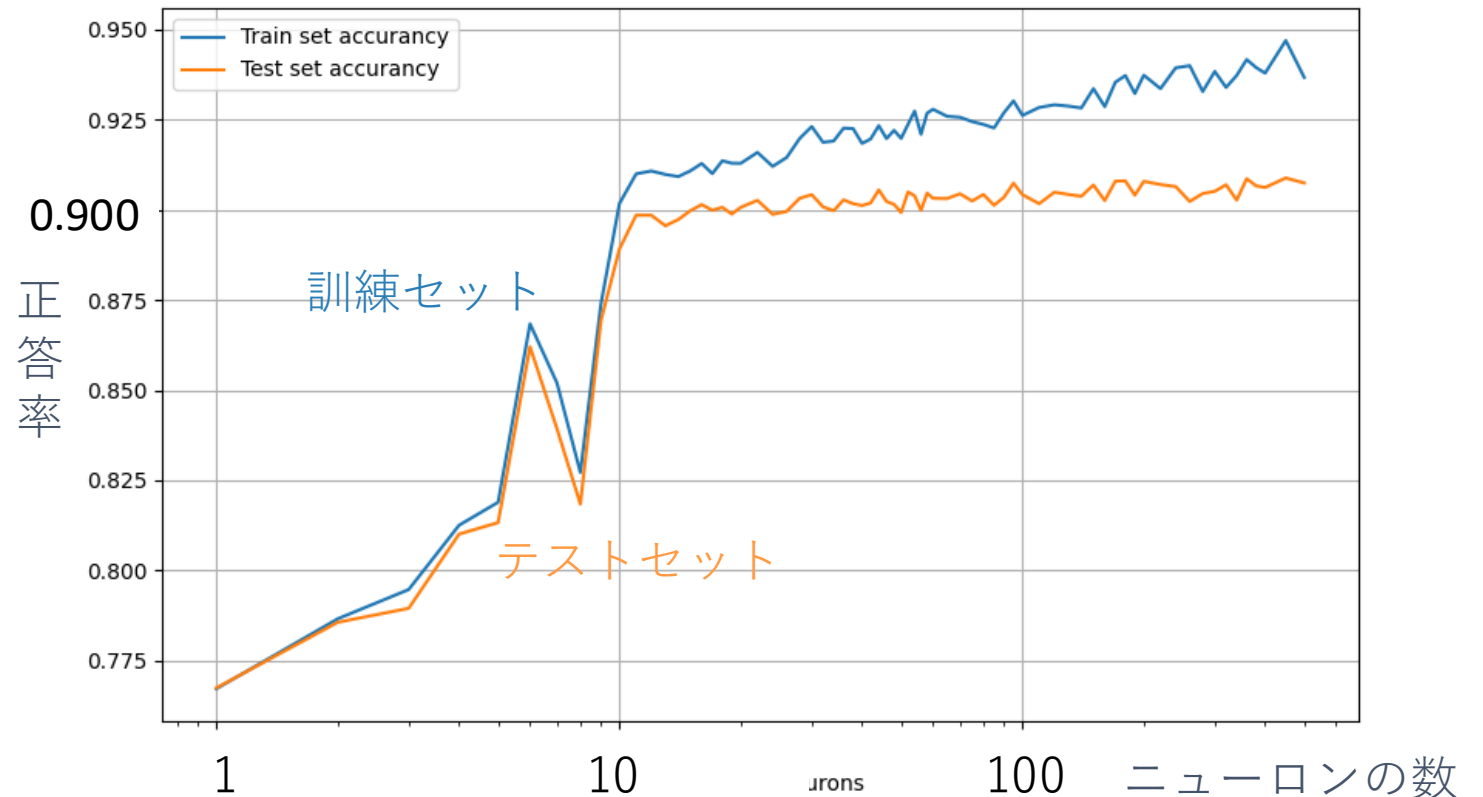
2 手法 特徴量



2 手法

モデルの設計

- 8割の天体で訓練
2割で性能を評価
- ハイパーパラメータは
(ニューロンの数など)
GridSearchにより決定
- 最適化アルゴリズム
adam
- 活性化関数
ReLU
- 損失関数
Cross Entropy関数
- 交差検証、層化抽出
StratifiedKFold

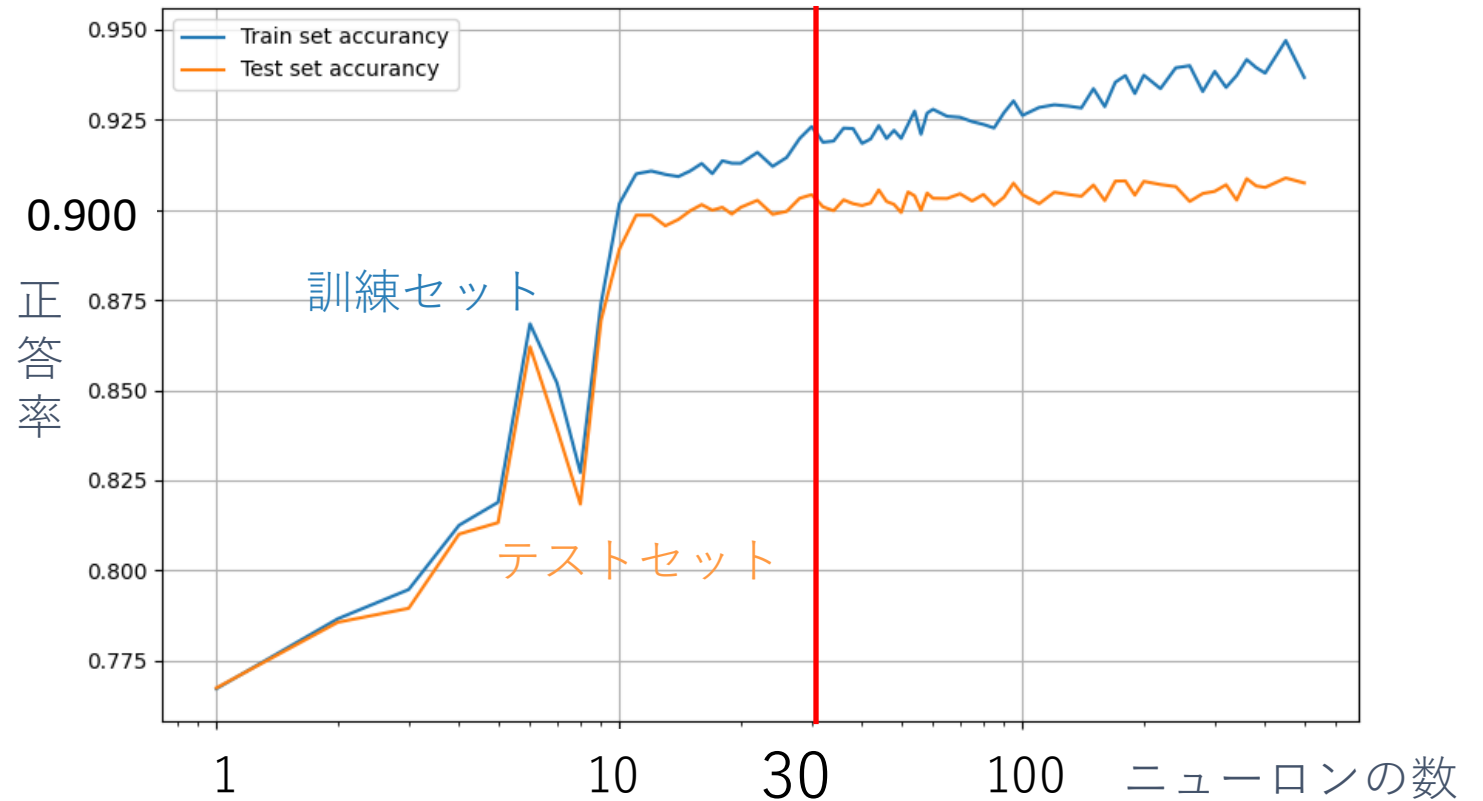


ニューロンの数の増加に伴う正答率の推移
今回は各層30個のニューロンの中間層を2層用いた

2 手法

モデルの設計

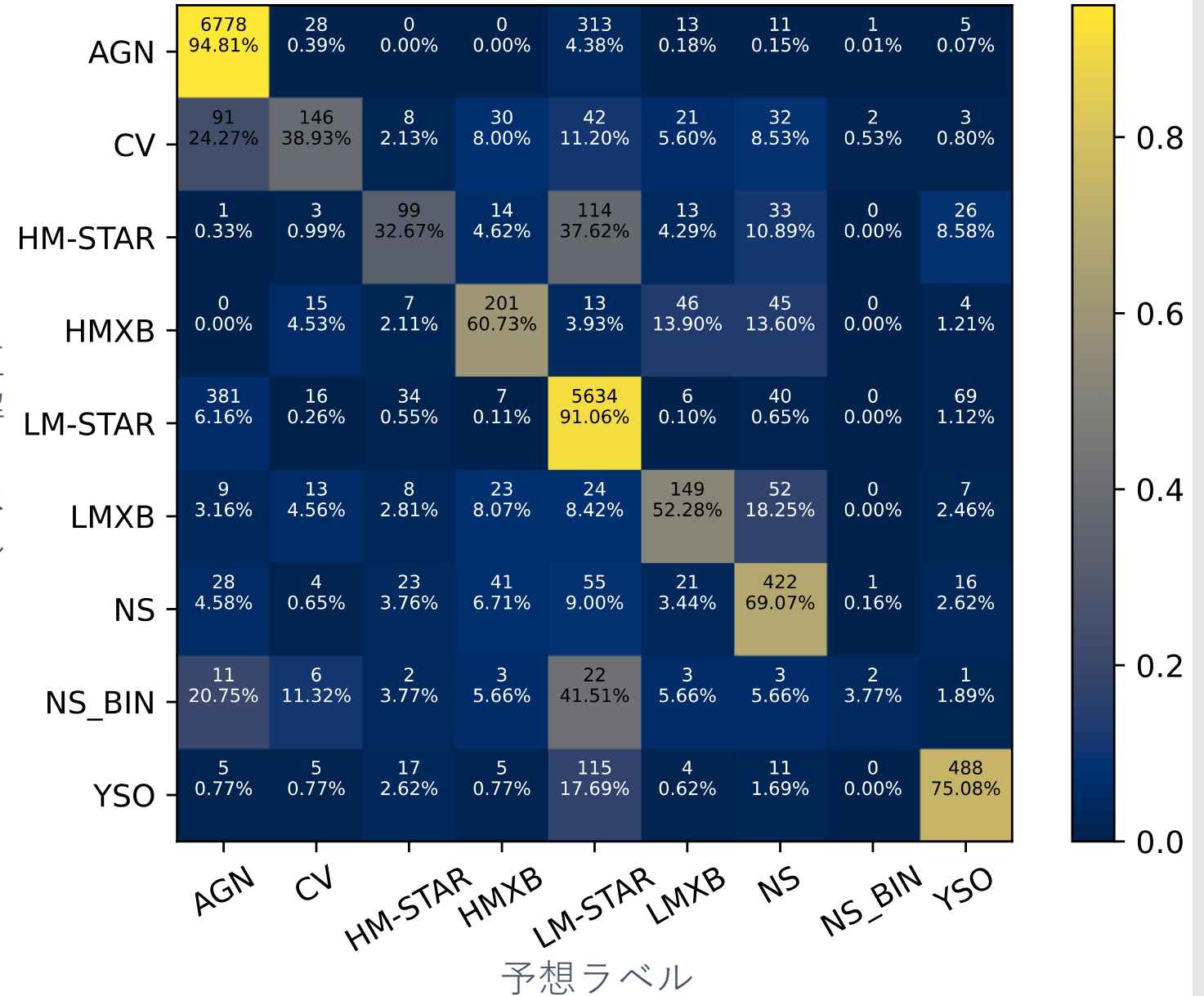
- 8割の天体で訓練
2割で性能を評価
- ハイパーパラメータは
(ニューロンの数など)
GridSearchにより決定
- 最適化アルゴリズム
adam
- 活性化関数
ReLU
- 損失関数
Cross Entropy関数
- 交差検証、層化抽出
StratifiedKFold



ニューロンの数の増加に伴う正答率の推移
今回は各層30個のニューロンの中間層を2層用いた

3 結果と考察

正
解
ラ
ベ
ル

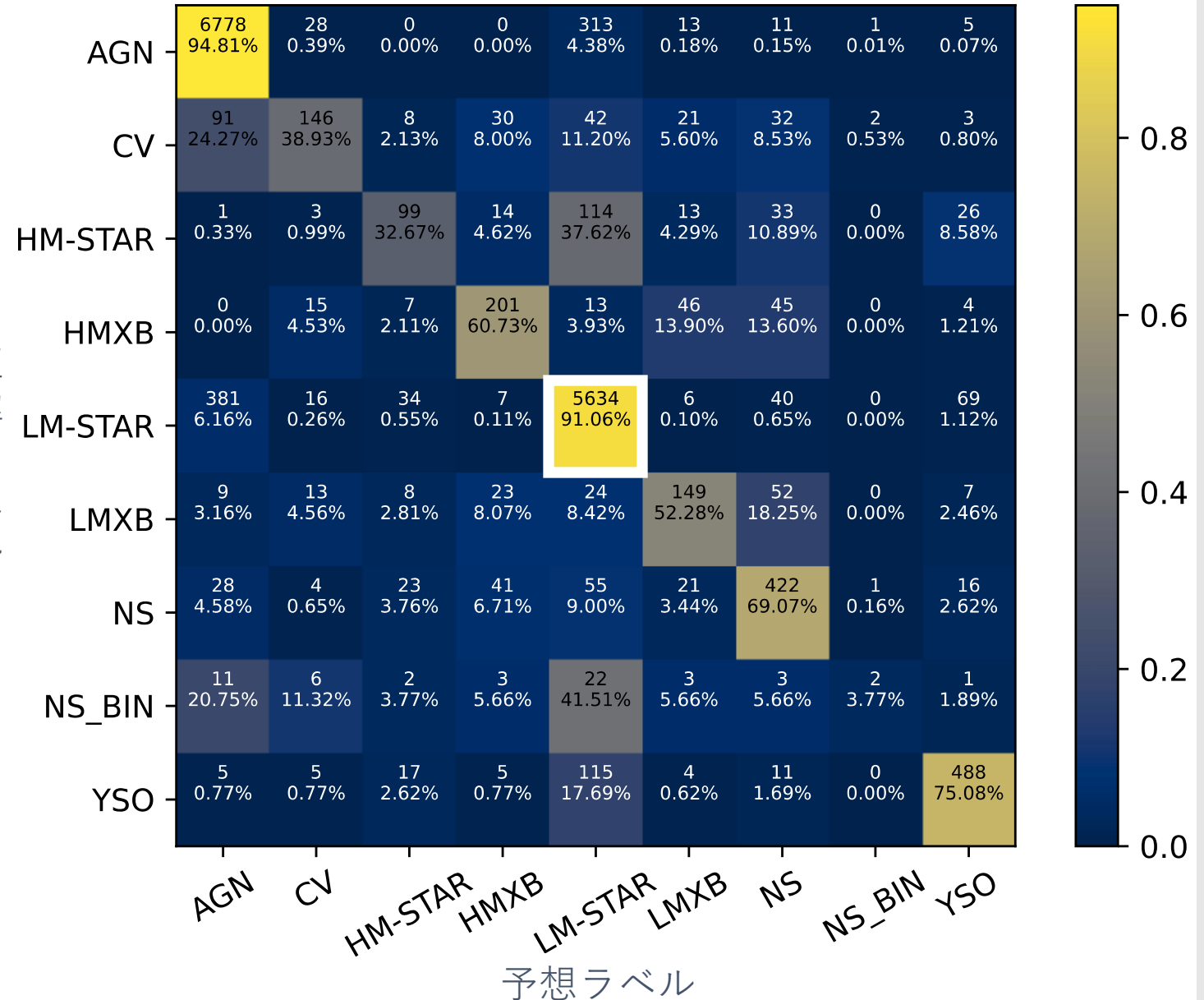


分類結果

3 結果と考察

例
正答 LM-STAR

正
解
ラ
ベ
ル



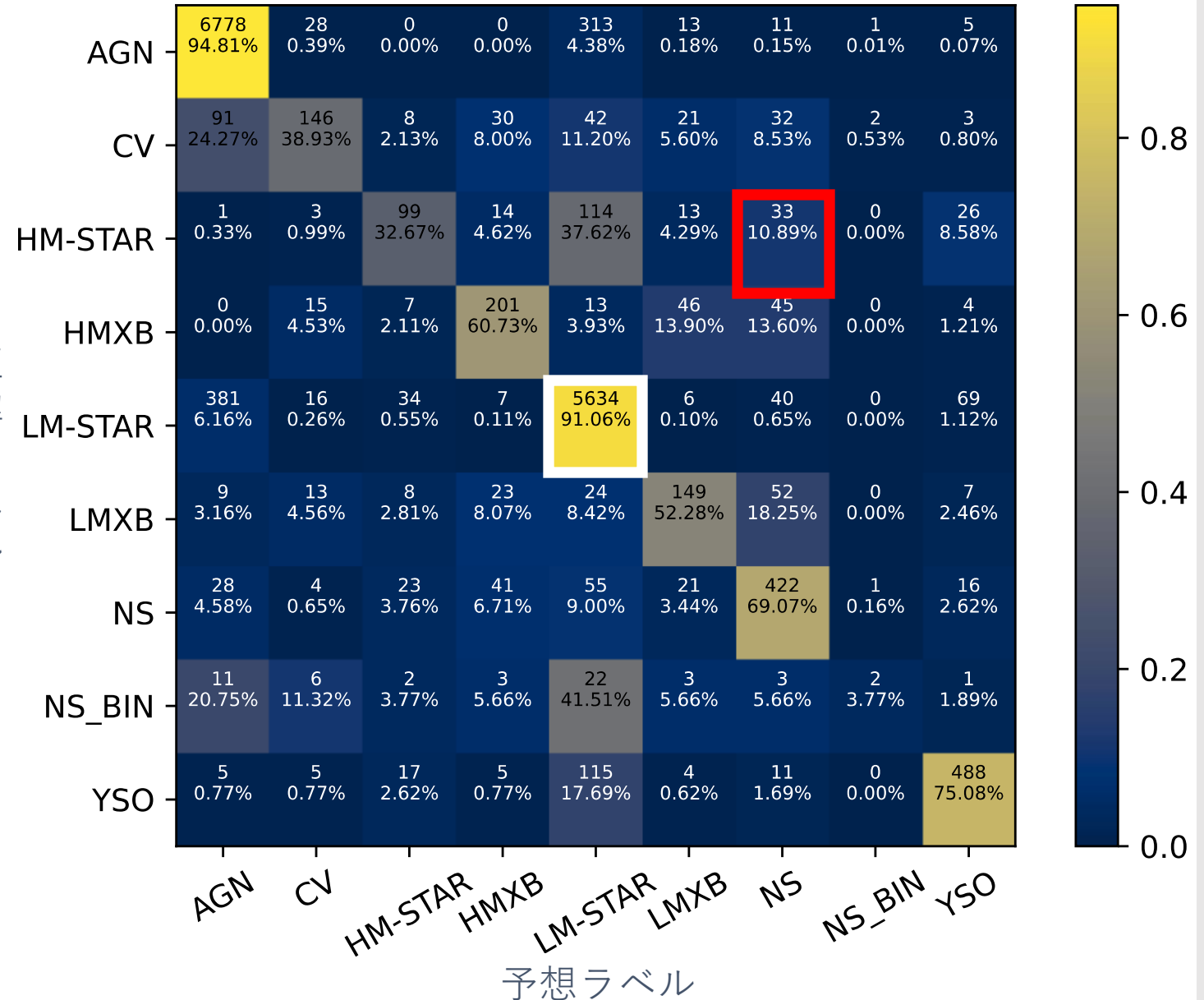
分類結果

3 結果と考察

例
正答 LM-STAR

誤答 HM-STAR → NS

正
解
ラ
ベ
ル



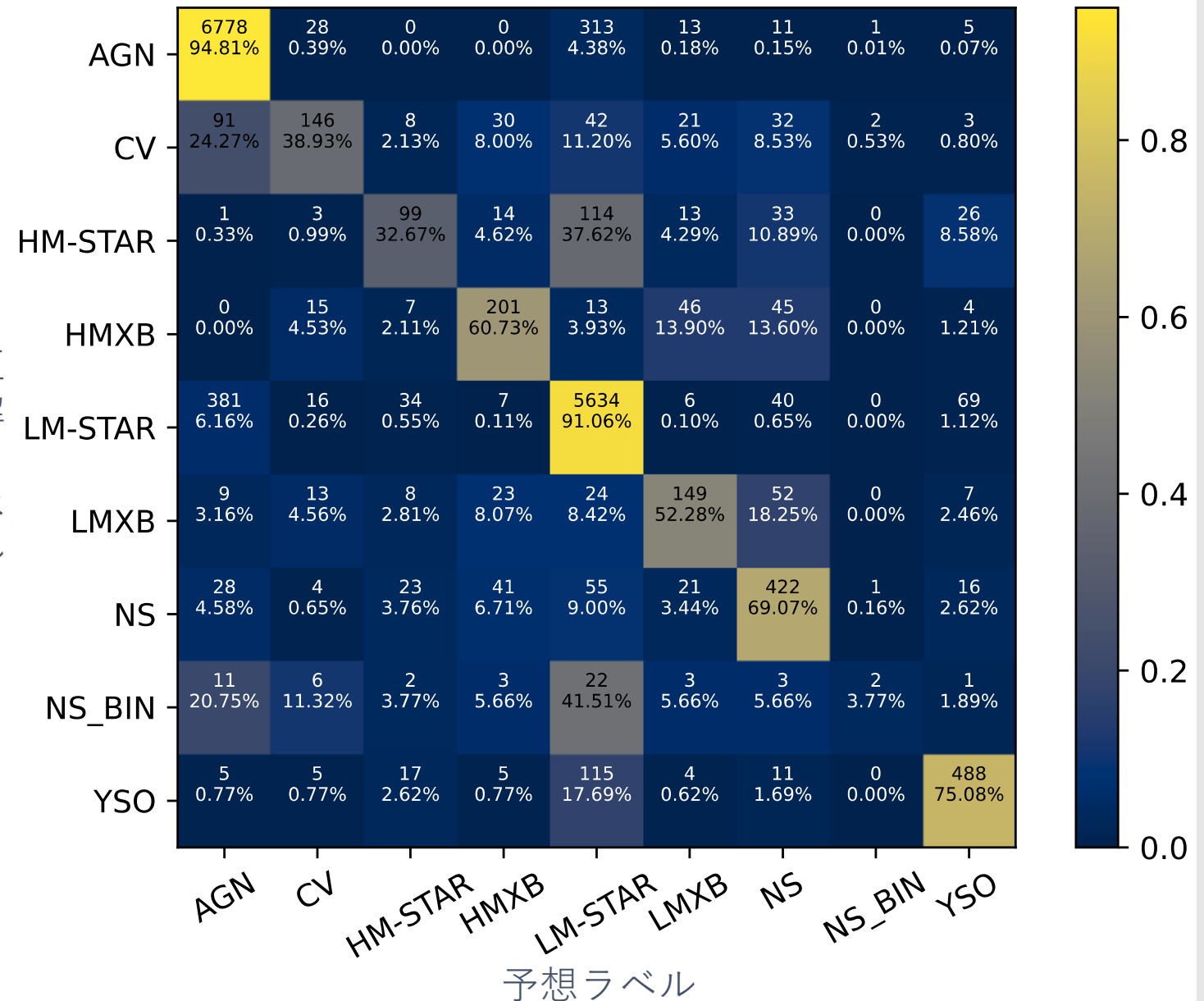
分類結果

3 結果と考察

正答率

87.3%

正
解
ラ
ベ
ル



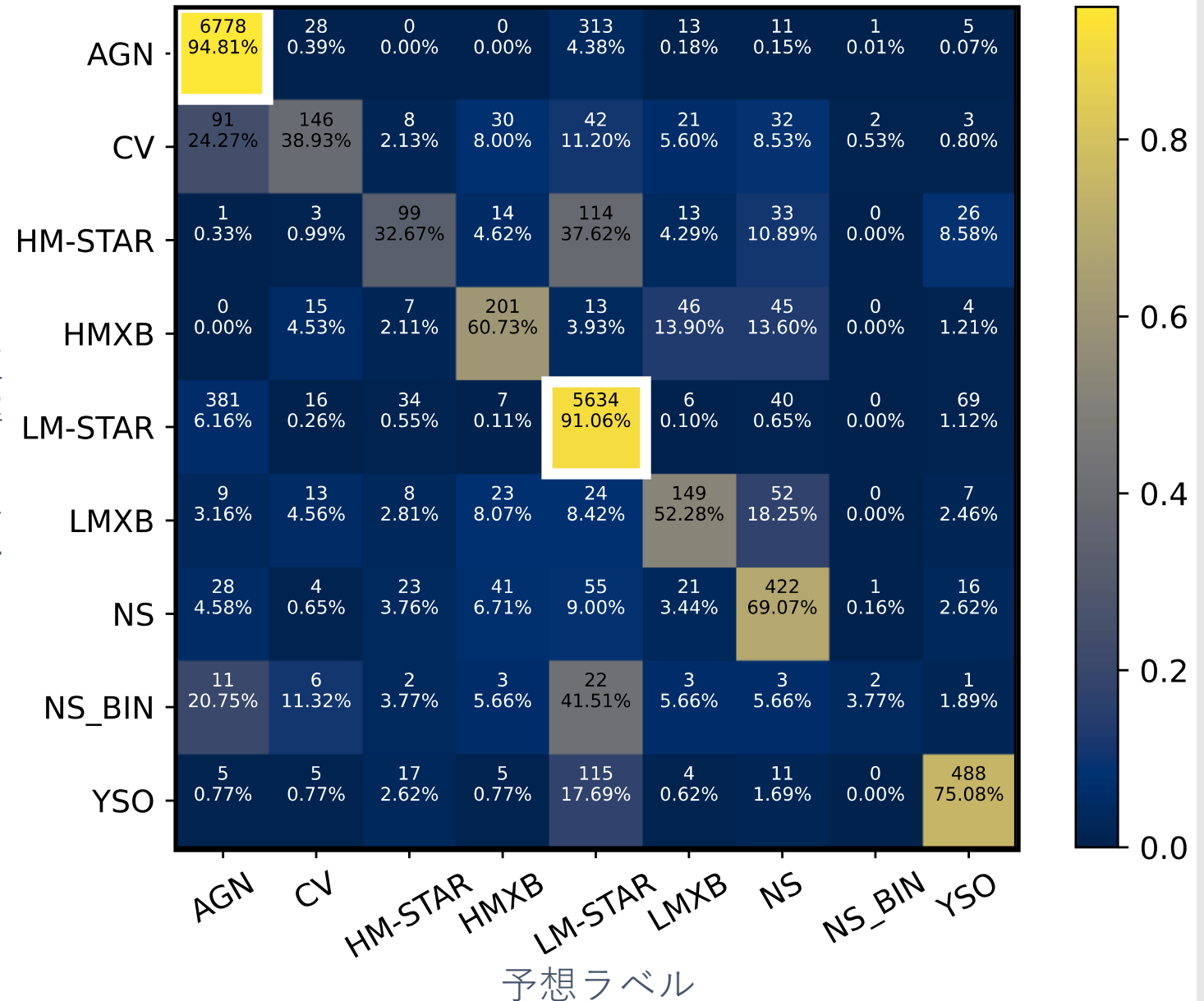
分類結果

3 結果と考察

正答率 87.3%

AGN・LM-STAR ○

正
解
ラ
ベ
ル



分類結果

3 結果と考察

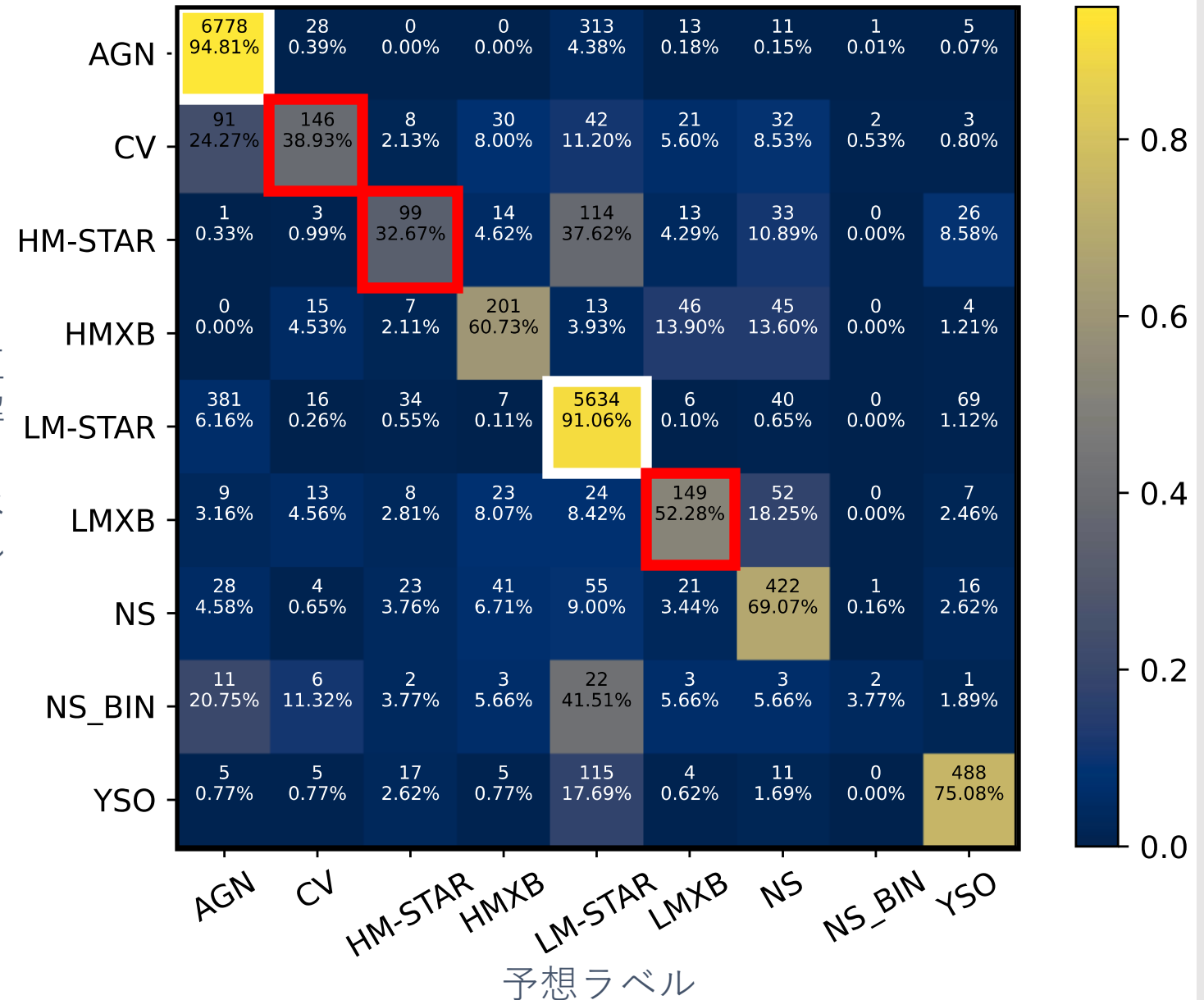
正答率 87.3%

AGN・LM-STAR ○

CV・HM-STAR
LMXB ▲

正解ラベル

分類結果



3 結果と考察

正答率 87.3%

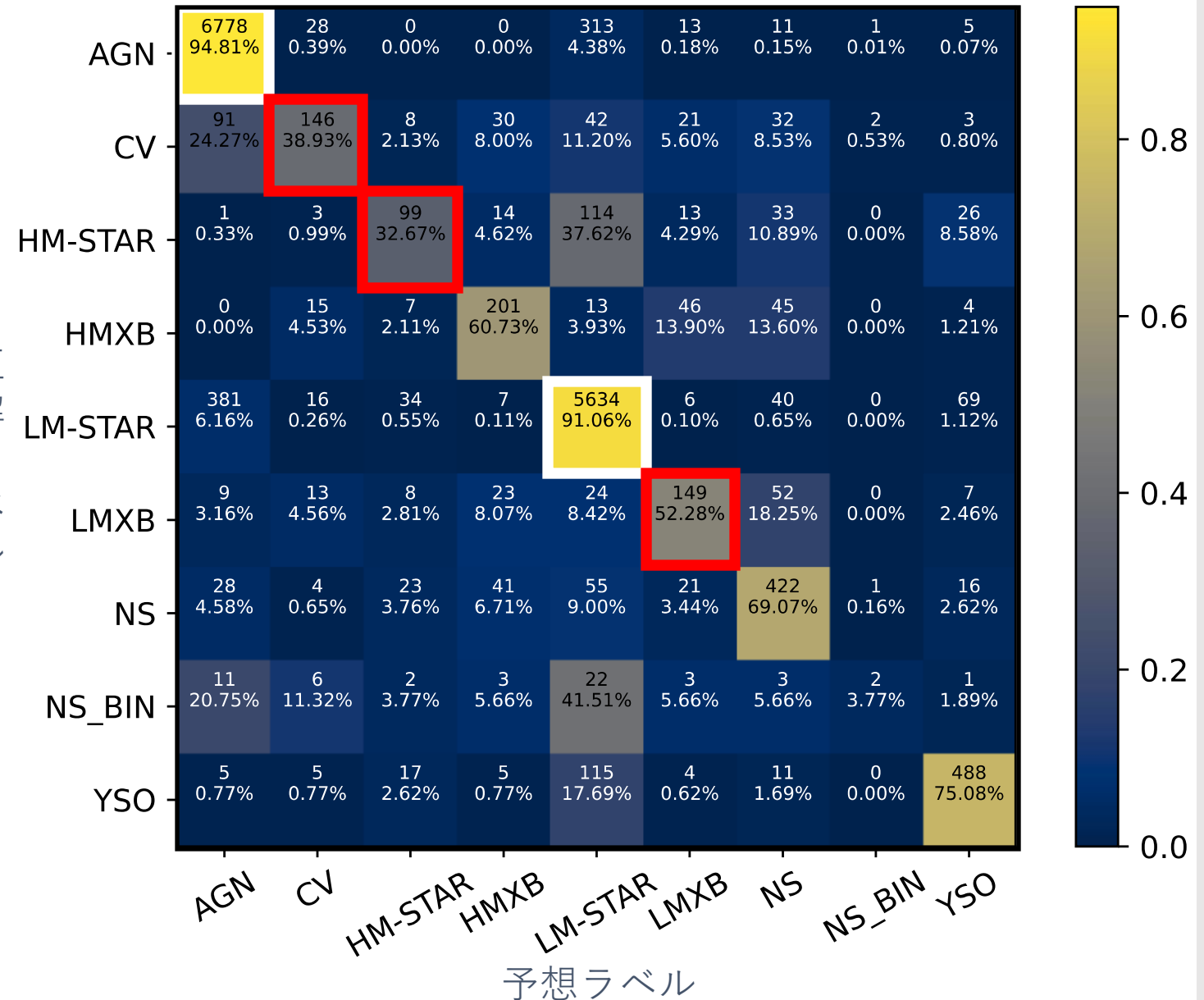
AGN・LM-STAR
→多数派クラス

CV・HM-STAR
LMXB
→少数派クラス



正
解
ラ
ベ
ル

分類結果



3 結果と考察

正答率 87.3%

AGN・LM-STAR
→多数派クラス

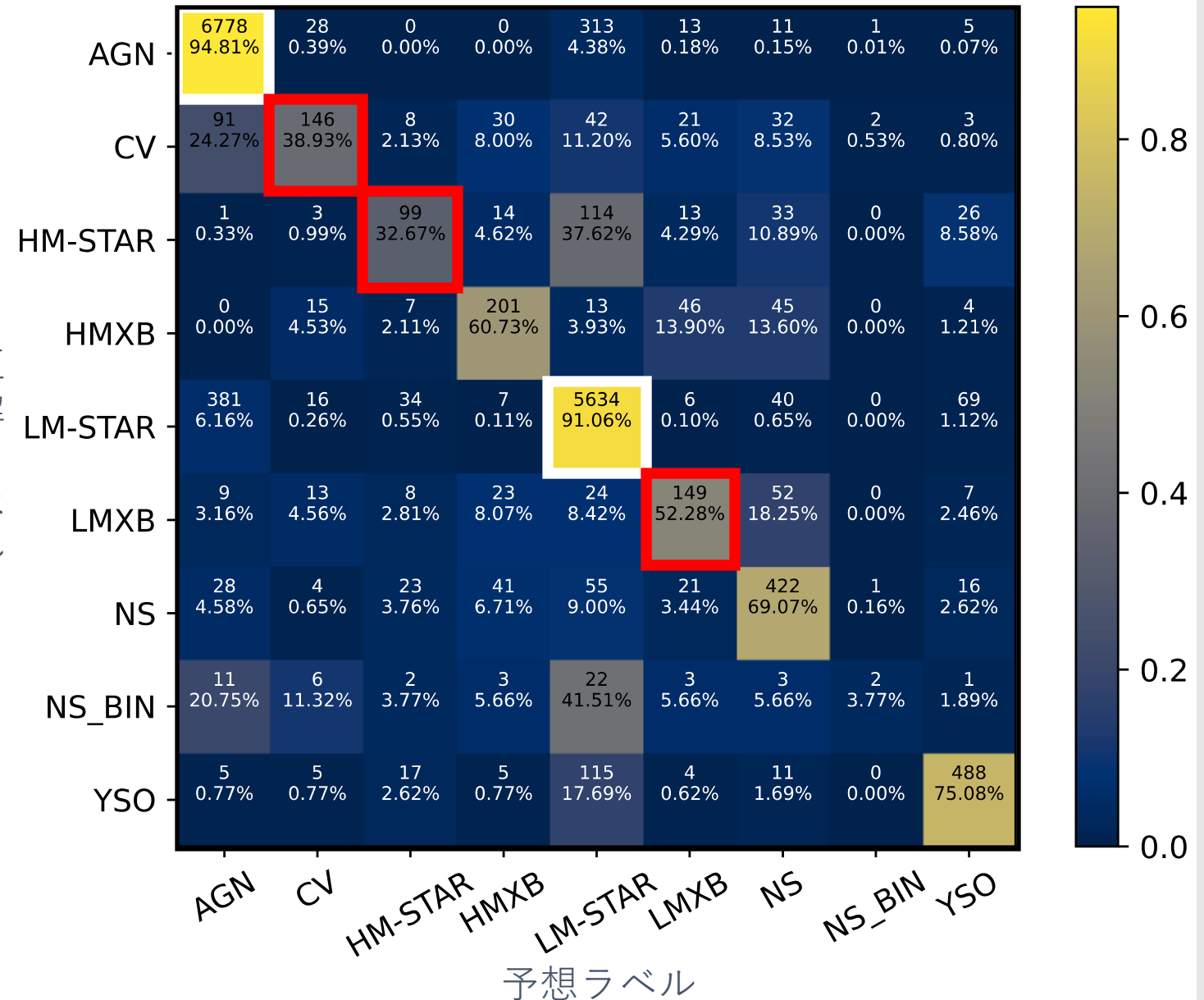
CV・HM-STAR
LMXB
→少数派クラス

多数派クラスをより重視?



正
解
ラ
ベ
ル

分類結果



3 結果と考察

正答率 87.3%

AGN・LM-STAR
→多数派クラス

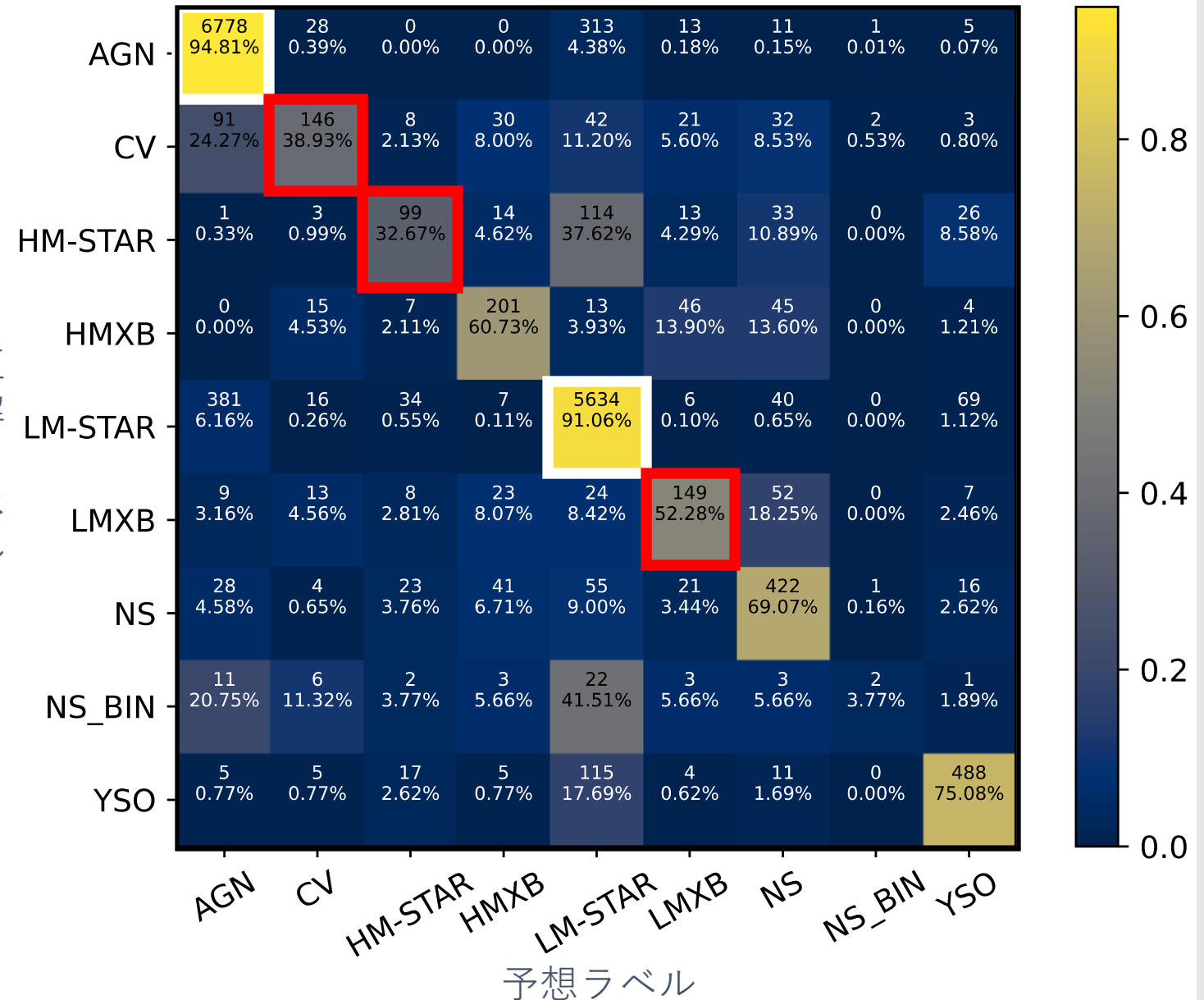
CV・HM-STAR
LMXB
→少数派クラス

多数派クラスをより重視?
↓
リサンプリング



正
解
ラ
ベ
ル

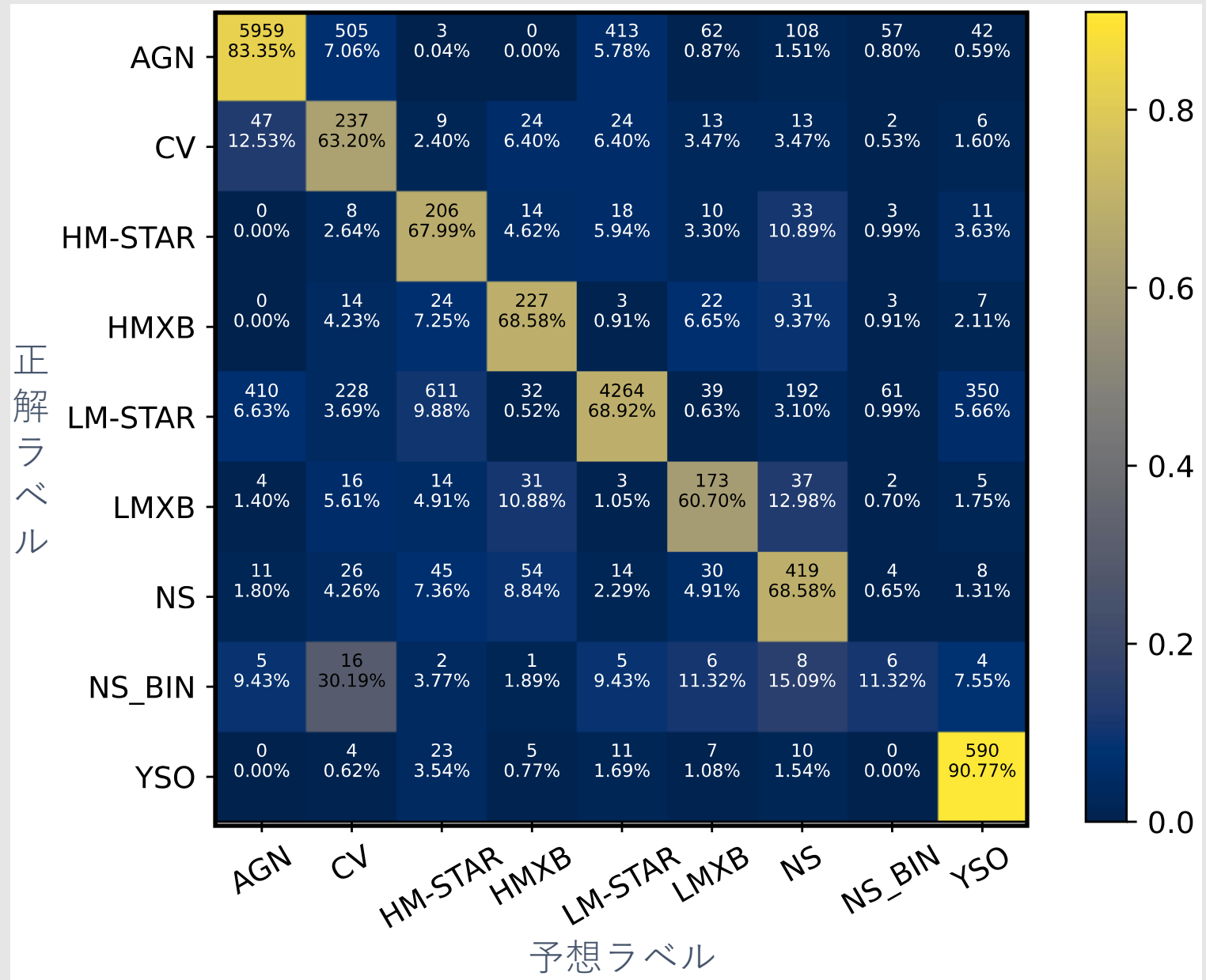
分類結果



3 結果と考察

正答率 75.8%
(-11.5%)

対角成分の正答率は上昇



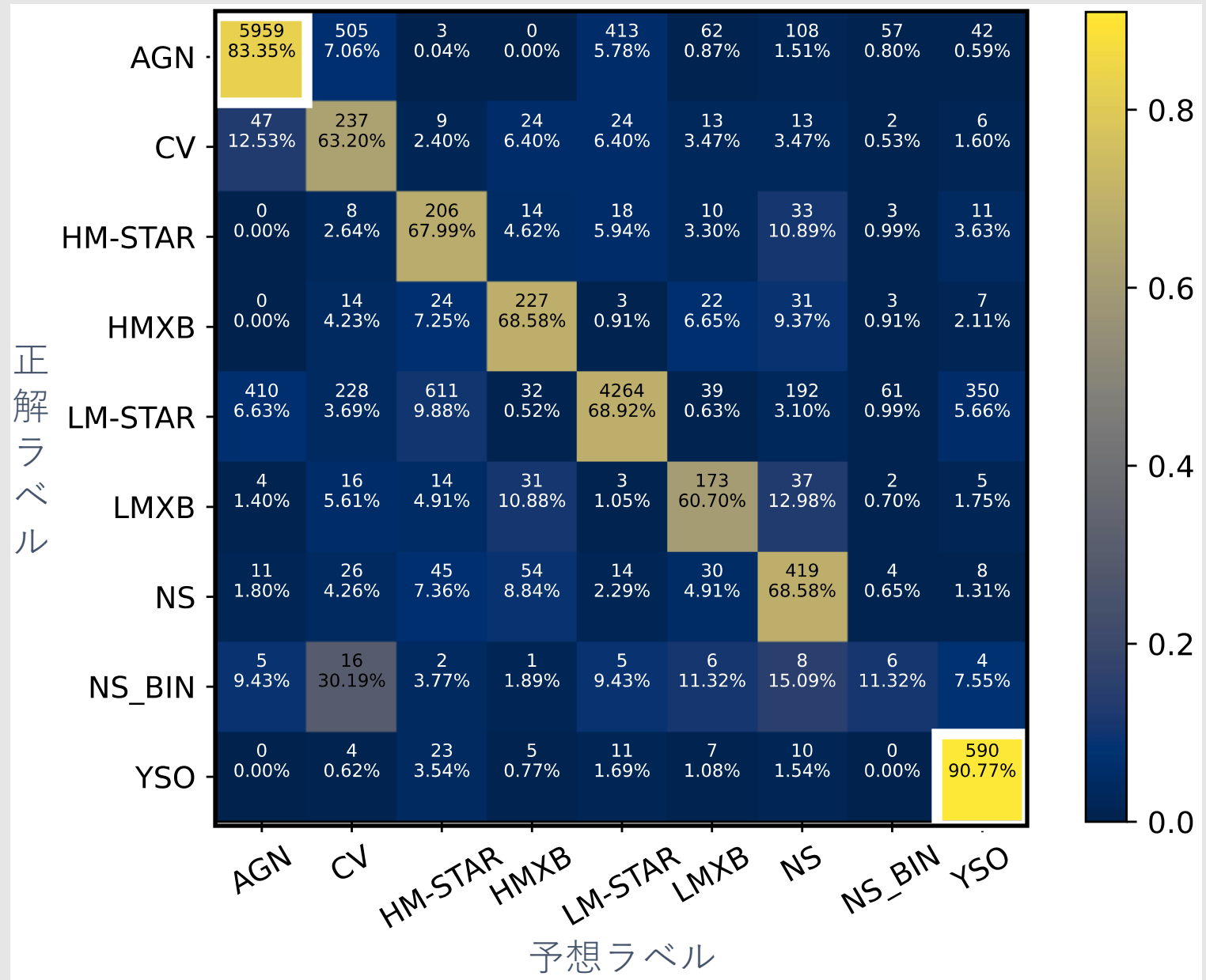
リサンプリングモデルの分類結果

3 結果と考察

正答率 75.8%
(-11.5%)

対角成分の正答率は上昇

AGN・YSO



リサンプリングモデルの分類結果

3 結果と考察

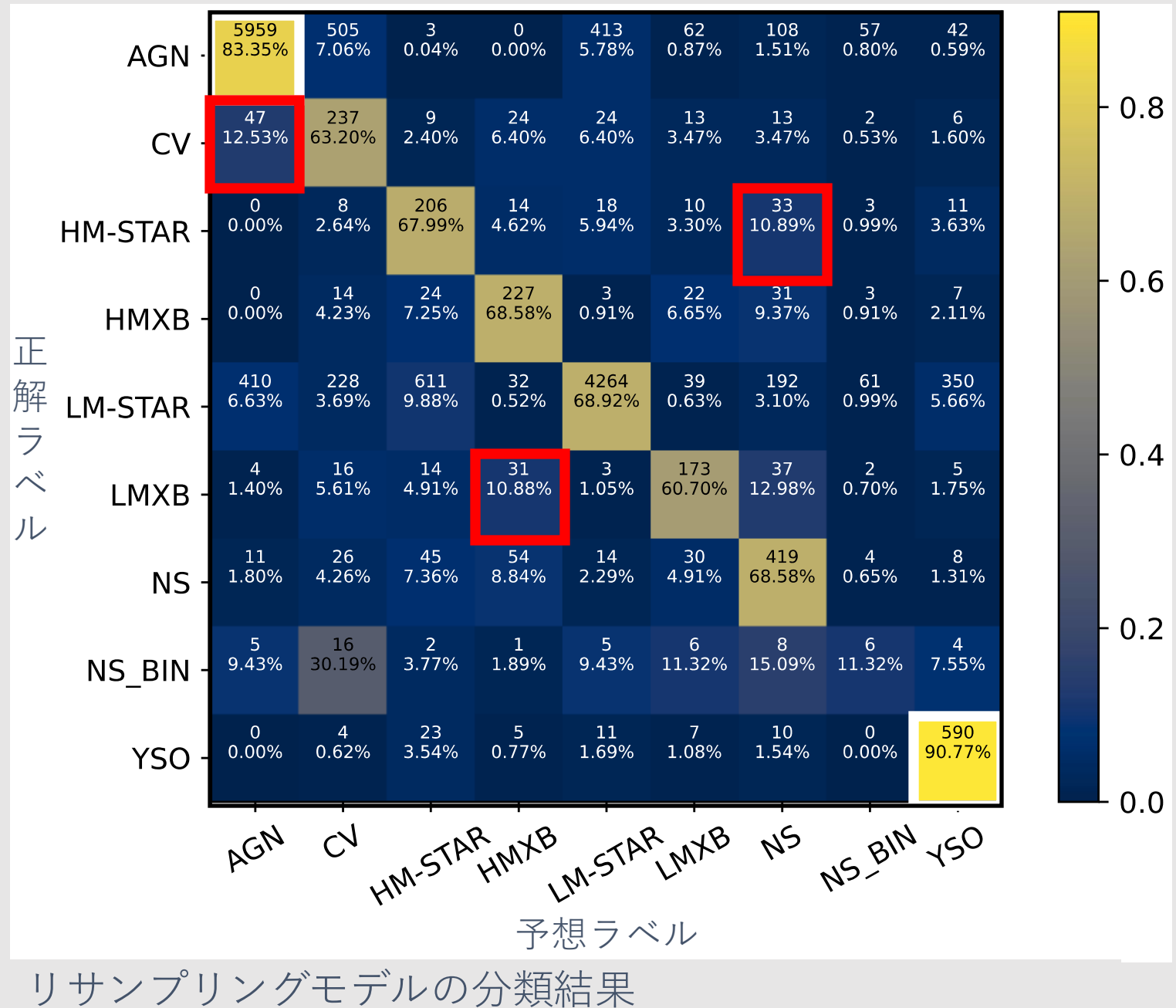
正答率 75.8%
(-11.5%)

対角成分の正答率は上昇

AGN・YSO

比較的頻度の高い誤答

CV → AGN
HM-STAR → NS
LMXB → HMXB
など



3 結果と考察

正答率 75.8%
(-11.5%)

対角成分の正答率は上昇

AGN・YSO 

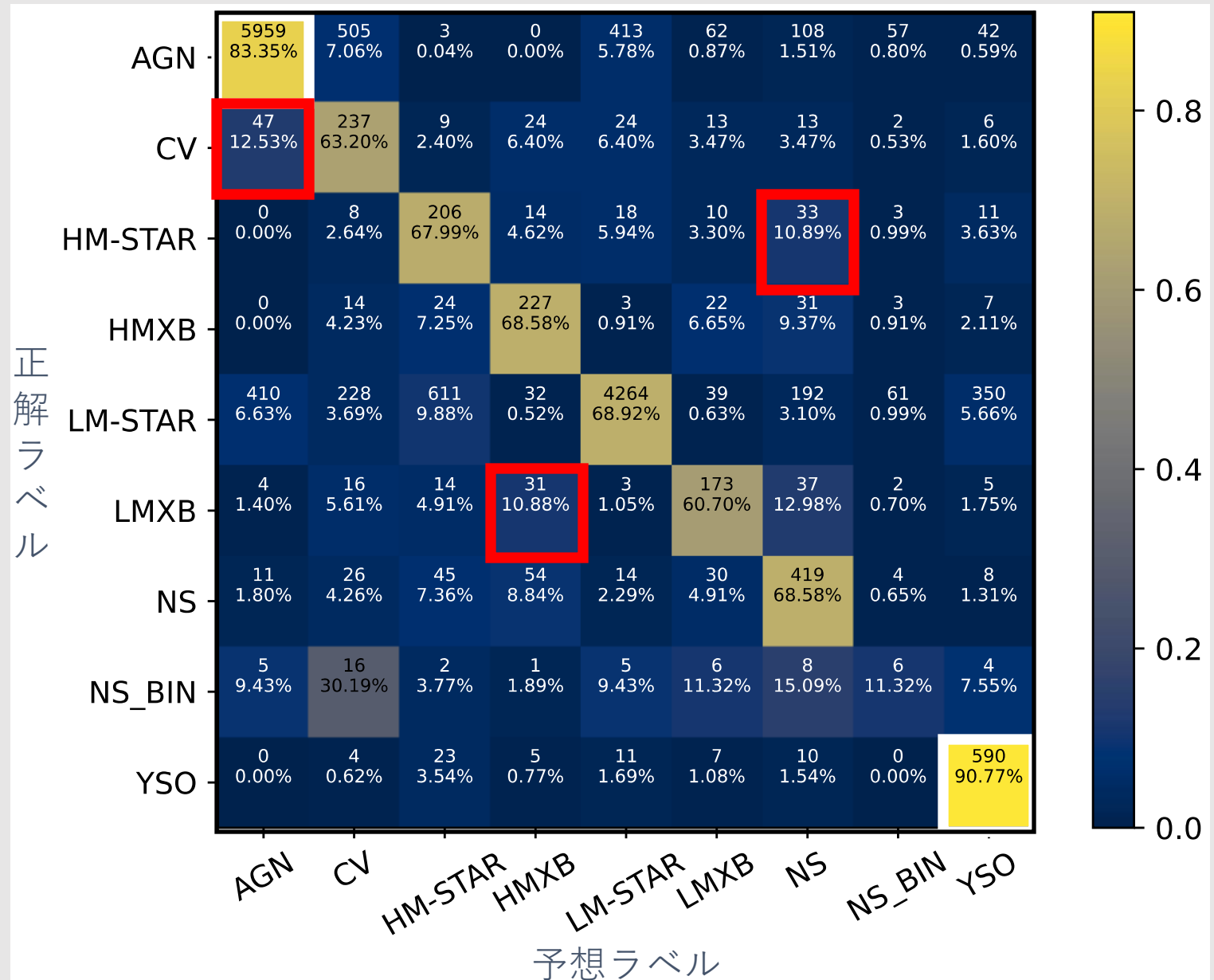
比較的頻度の高い誤答

CV → AGN

HM-STAR → NS

LMXB → HMXB

など

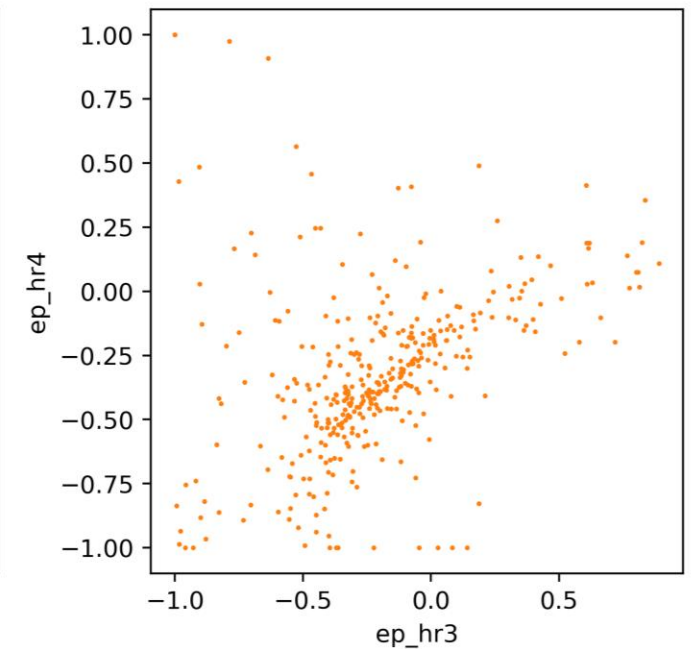
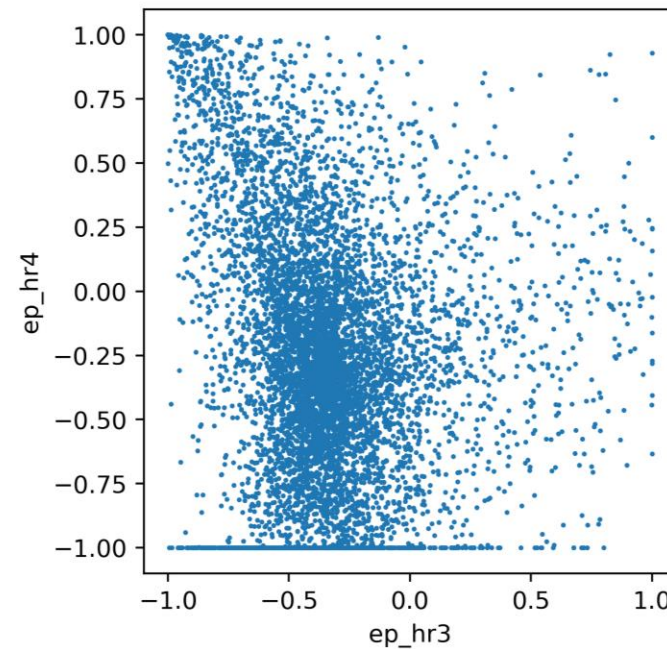
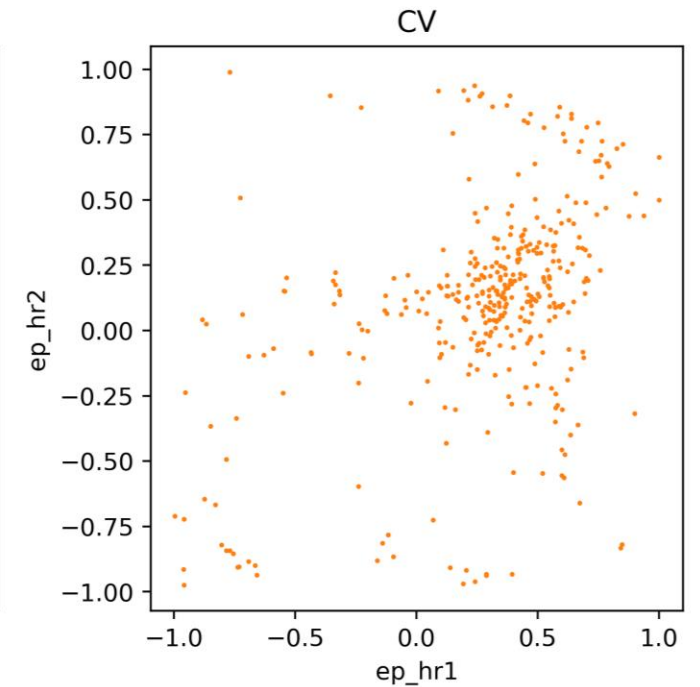
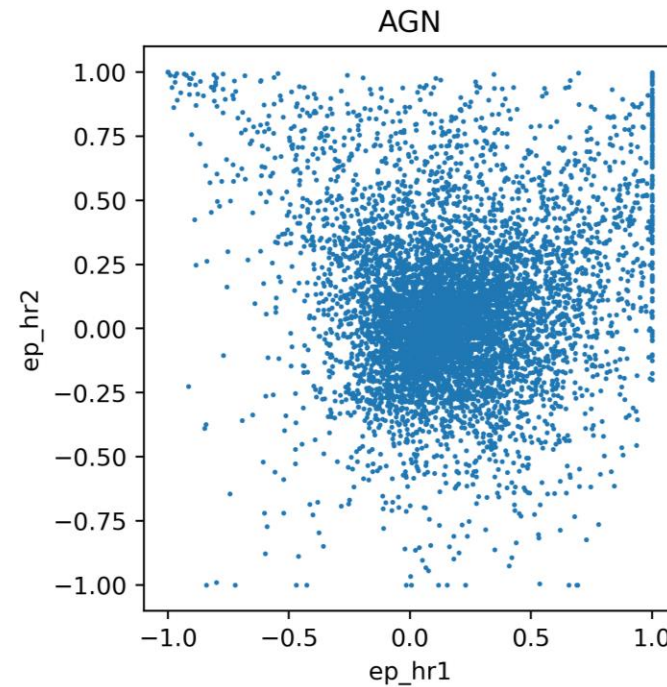


リサンプリングモデルの分類結果

3 結果と考察

CVの分布が
AGNの分布に含まれている

→区別が付きづらい



AGNとCVの硬度比をプロットしたもの

4 まとめ

- ・ニューラルネットワークによるX線天体の分類

1 天体の8割を訓練セットとした分類



2 リサンプリングモデルでの分類

正答率 75.8%

- ・今後の課題

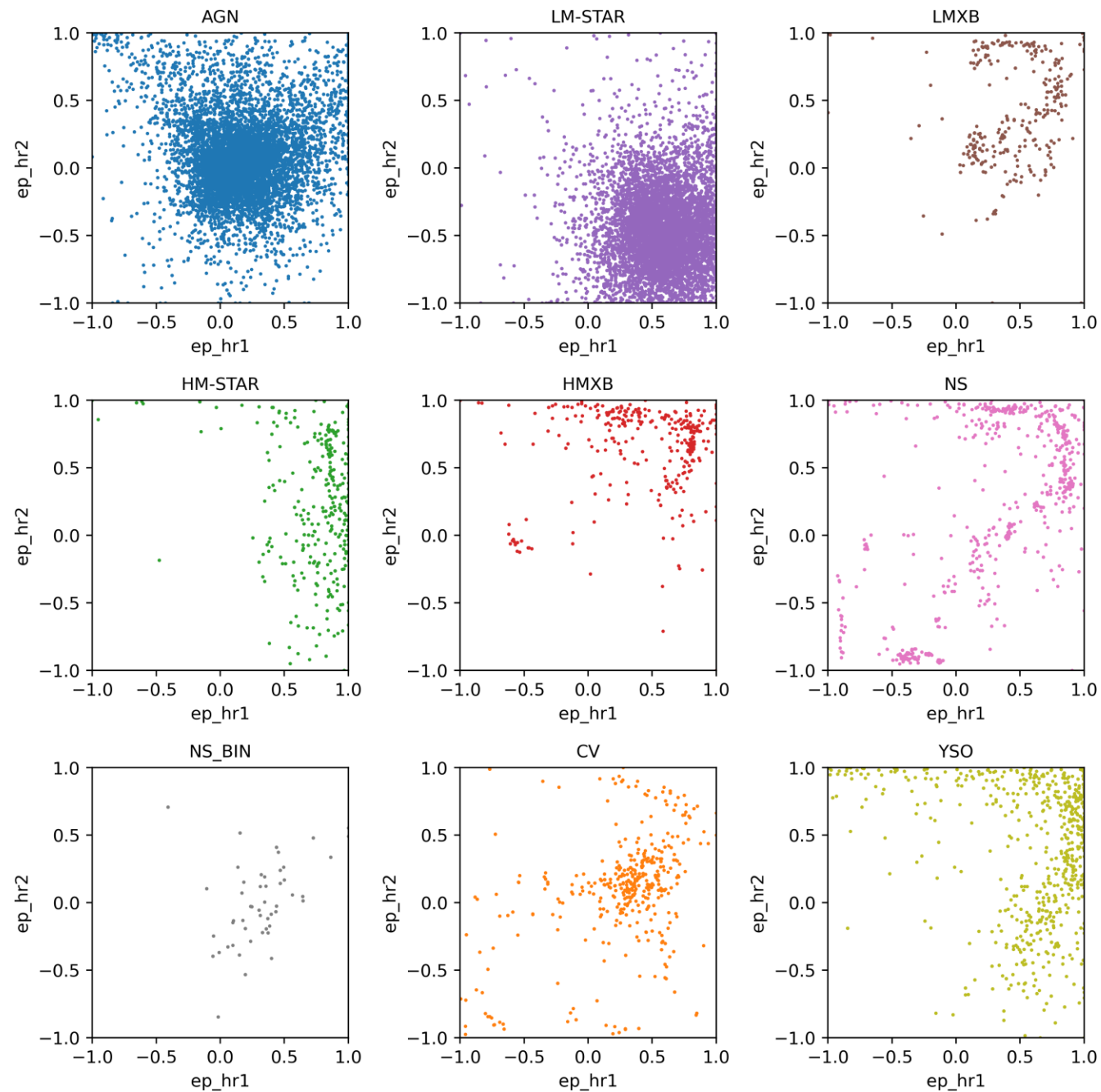
有用なデータの追加

→時系列データ(ライトカーブなど)、X線以外の波長など

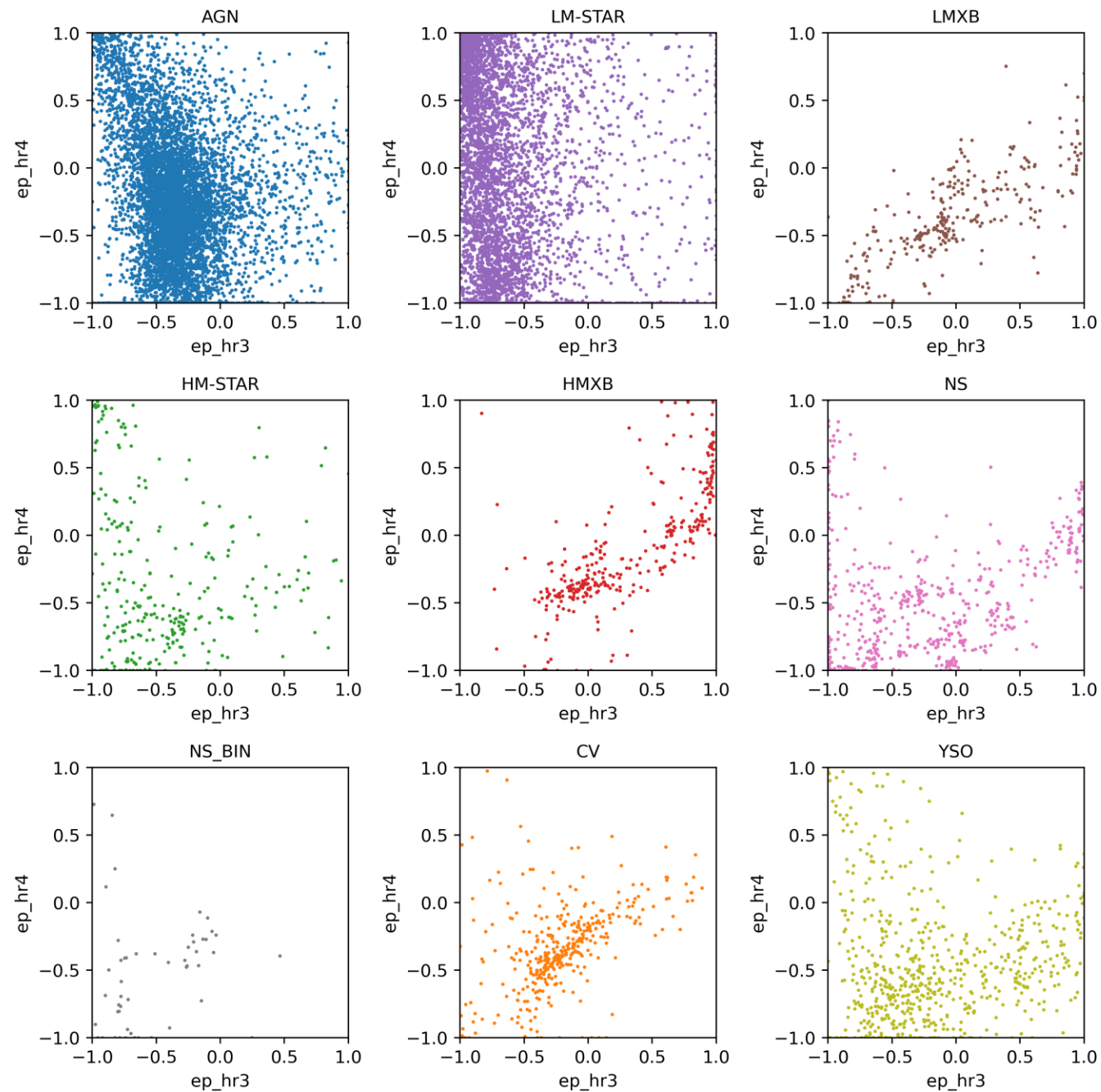
今回の教師データに銀河団などは含まれていない

補足資料

HR_1 vs HR_2



HR_3 vs HR_4



欠損の割合

	AG N	715 3	CV	373	HM- STA R	299	HMX B	322	LM- STAR	6231	LMX B	292	NS	614	NS_ BIN	50	YS O	650
ep_8_rate	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%
ep_extent	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%
ep_hr1	6	0%	1	0%	1	0%	2	1%	1	0%	0	0%	4	1%	0	0%	10	2%
ep_hr2	1	0%	0	0%	0	0%	0	0%	0	0%	1	0%	0	0%	0	0%	0	0%
ep_hr3	2	0%	0	0%	0	0%	0	0%	33	1%	0	0%	0	0%	0	0%	0	0%
ep_hr4	28	0%	1	0%	2	1%	0	0%	306	5%	1	0%	3	0%	0	0%	6	1%
ep_chi2 prob	442 8	62%	35	9%	145	48%	27	8%	4167	67%	20	7%	58	9%	17	34%	325	50%
pn_fvar	616 1	86%	10 2	27%	240	80%	112	35%	5318	85%	170	58%	337	55%	40	80%	491	76%

表1:使用した特徴量の欠損値のクラスごとの数と割合

モデルの改良 リサンプリング

分類ごとの天体の数を調整しデータ間の不均衡を解消する
用いた手法 : SMOTE (Synthetic Minority Over-sampling Technique)
: ランダムアンダーサンプリング

クラス名	天体数	訓練セットの天体数	リサンプリング後の天体数 (訓練セット)
AGN	7159	5720	500
CV	375	300	500
HM-STAR	303	242	500
HMXB	331	265	500
LM-STAR	6187	4929	500
LMXB	285	228	500
NS	611	489	500
NS_BIN	53	43	80
YSO	650	520	520

表1 : リサンプリング前後の天体数

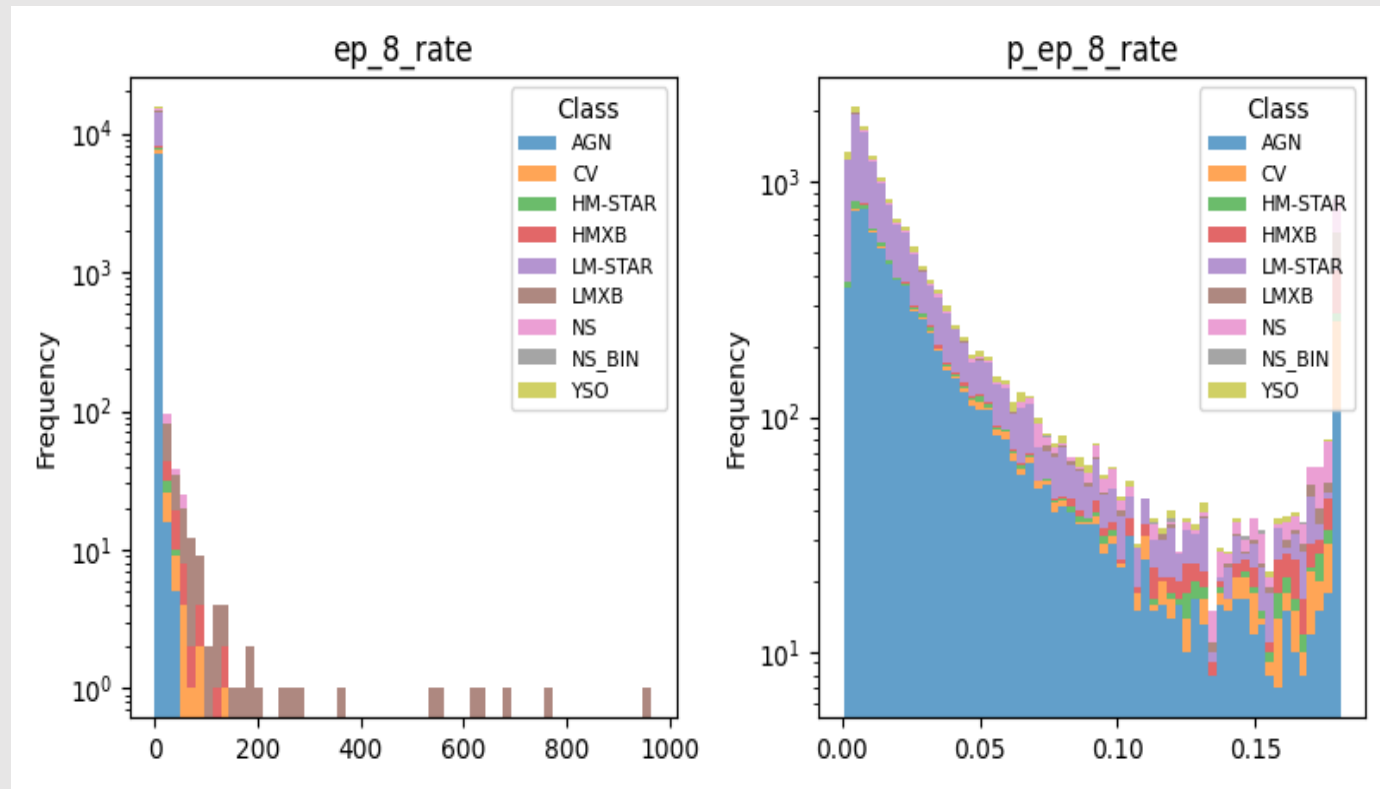
スケーリング PowerTransformer (Yeo-Johnson変換)

データを正規分布に近づける方法

$$\begin{aligned}x'(\lambda) &= \frac{(x+1)^\lambda - 1}{\lambda} & (\lambda \neq 0, x_i \geq 0) \\&= -\frac{(x+1)^{2-\lambda} - 1}{2-\lambda} & (\lambda \neq 2, x_i < 0) \\&= \text{sign}(x) \log(x+1) & (\lambda = 0, x_i \geq 0)\end{aligned}$$

$$\begin{aligned}\text{sign}(x) &= 1 \quad (x > 0) \\&= 0 \quad (x = 0) \\&= -1 \quad (x < 0)\end{aligned}$$

λ は変換後の分布が正規分布
に近づくよう最尤法で決定される
パラメータ

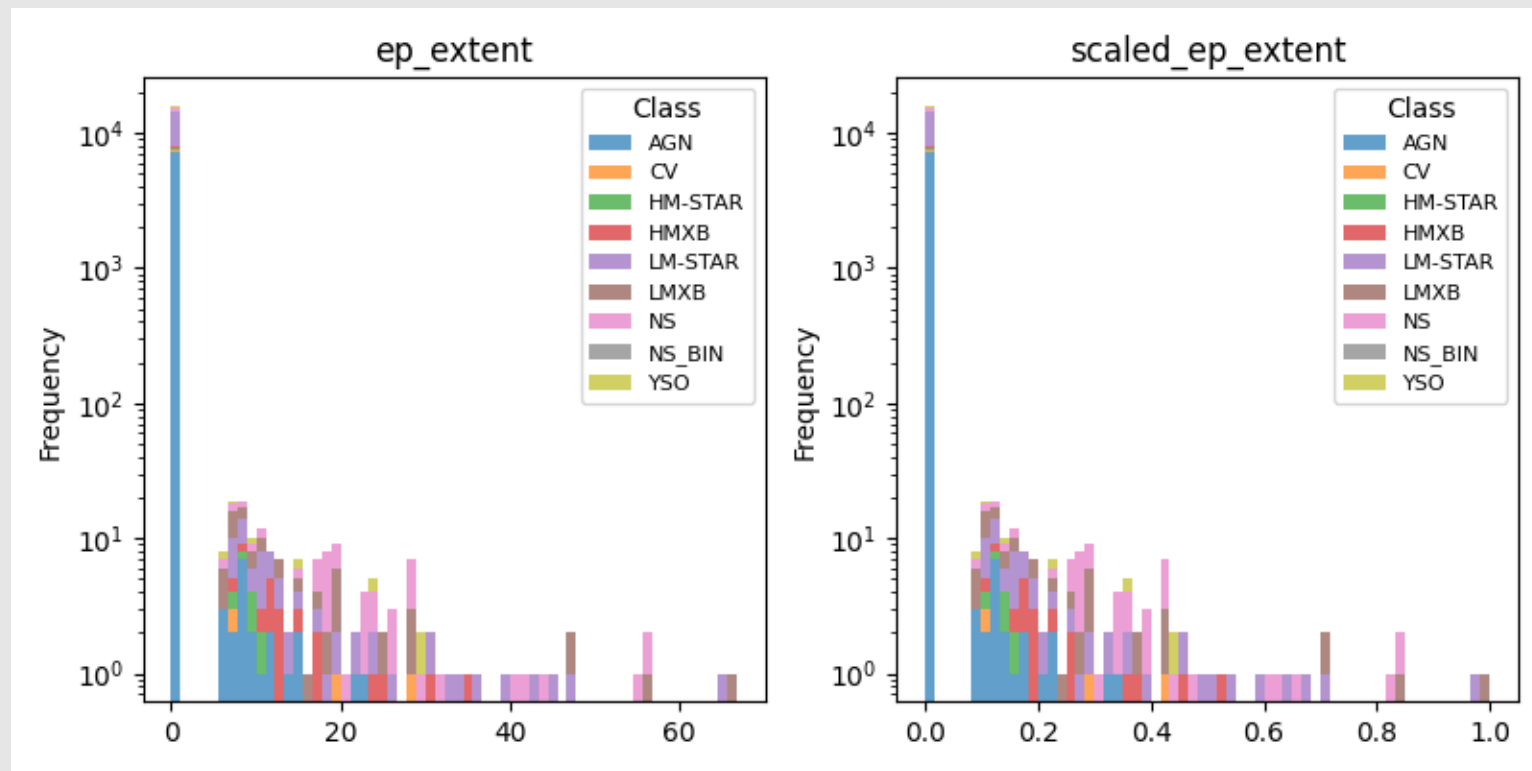


スケーリング前(左)と後(右)のヒストグラム

スケーリング MinMaxScaler 最小最大スケーリング

$$x_i' = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

[0, 1]の範囲に線形変換
正規化とも呼ぶ



スケーリング前(左)とスケーリング後(右)

損失関数 Cross Entropy

$$H(p, q) = - \sum_x p(x) \log(q(x))$$

p : 真の分布確率

q : 推定した分布確率

真の分布と推定した分布の違いを比べるために用いられる

pとqが近いほどHは小さくなる

勾配降下法

最適化アルゴリズムの一つ

$$w_{i+1} = w_i - \eta \frac{\partial E}{\partial w}$$

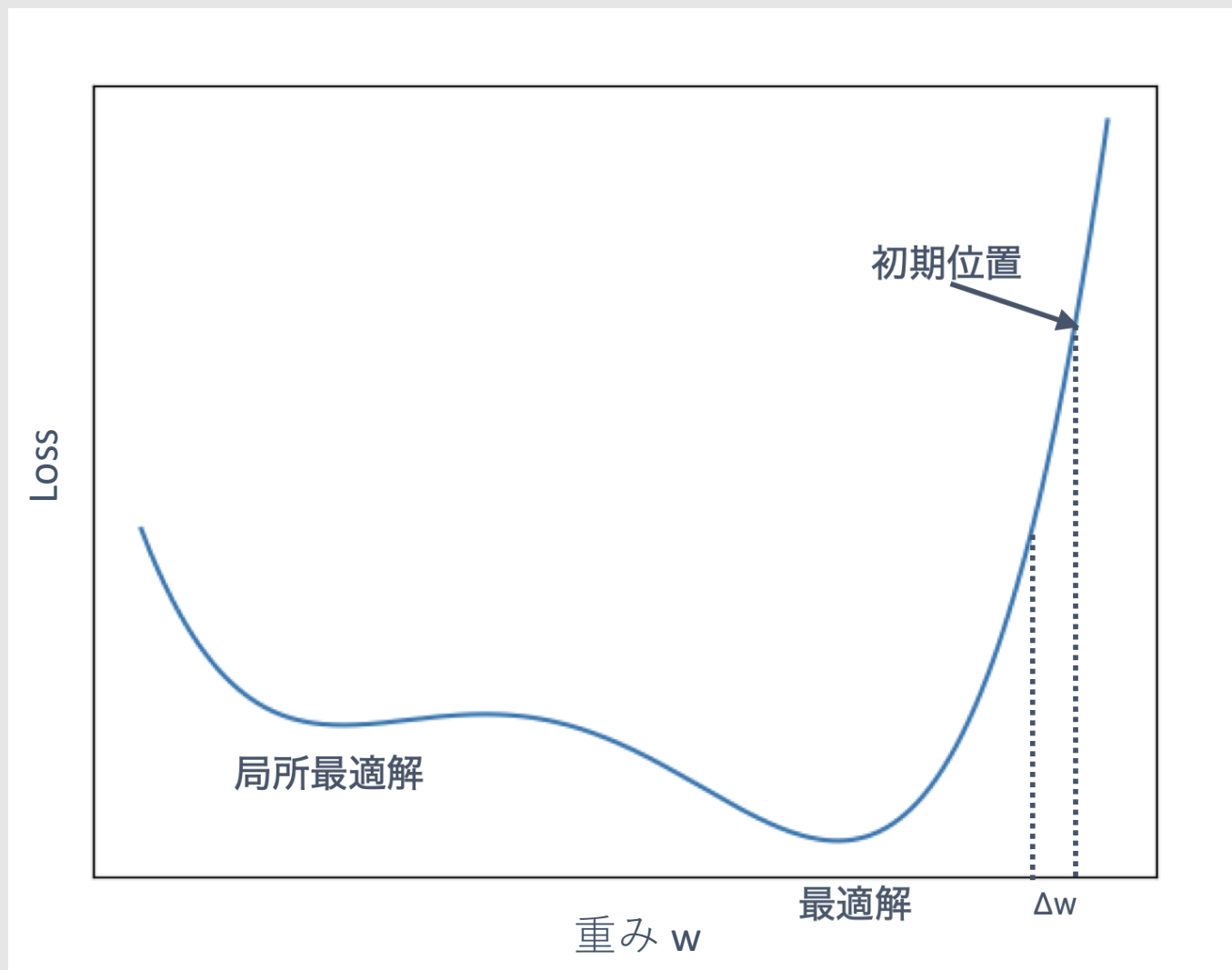
w : 重み

η : 学習率

E : 損失関数

その点での勾配を計算し
次の重みを計算する

重みが変わらなくなった点で
決定



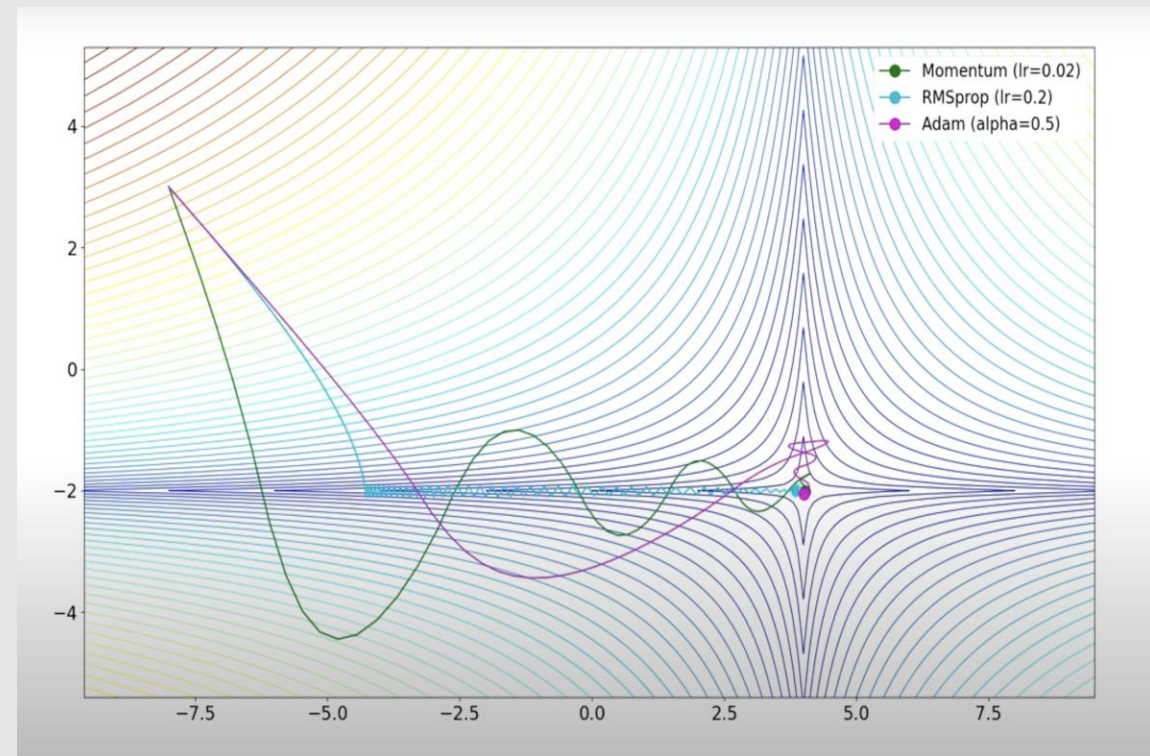
勾配降下法

最適化の方法 Solver : Adam

Adam(adaptive moment estimation)

Momentum SGDとRMSpropの組み合わせ
過去の勾配の指数関数的減衰平均と
勾配の二乗の指数関数的減衰平均を参照

$$\begin{aligned} m_t &\leftarrow \beta_1 m_{t-1} + (1 - \beta_1) \Delta w_t \\ v_t &\leftarrow \beta_2 v_{t-1} + (1 - \beta_2) (\Delta w_t)^2 \\ \hat{m}_t &= \frac{m_t}{1 - \beta_1^t} \\ \hat{v}_t &= \frac{v_t}{1 - \beta_2^t} \\ w_{t+1} &\leftarrow w_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \epsilon}} \end{aligned}$$



Momentum SGDとRMSpropとAdamの最適化の過程

誤差逆伝播法

損失を最小化する最適化を、順方向である入力層からではなく出力層から(逆方向)から行う
誤差が逆方向に伝播していく様子から名付けられた

逆方向から行うメリット

連鎖律を用いて重みパラメータの勾配を求める

その時後ろのニューロンのパラメータも参照する

前からパラメータの更新を行うと後ろのパラメータも変わるため

何度も同じ計算をしなくてはならない

逆方向から計算を行うため同じ計算をする必要がない

→計算時間の短縮

活性化関数 ReLU

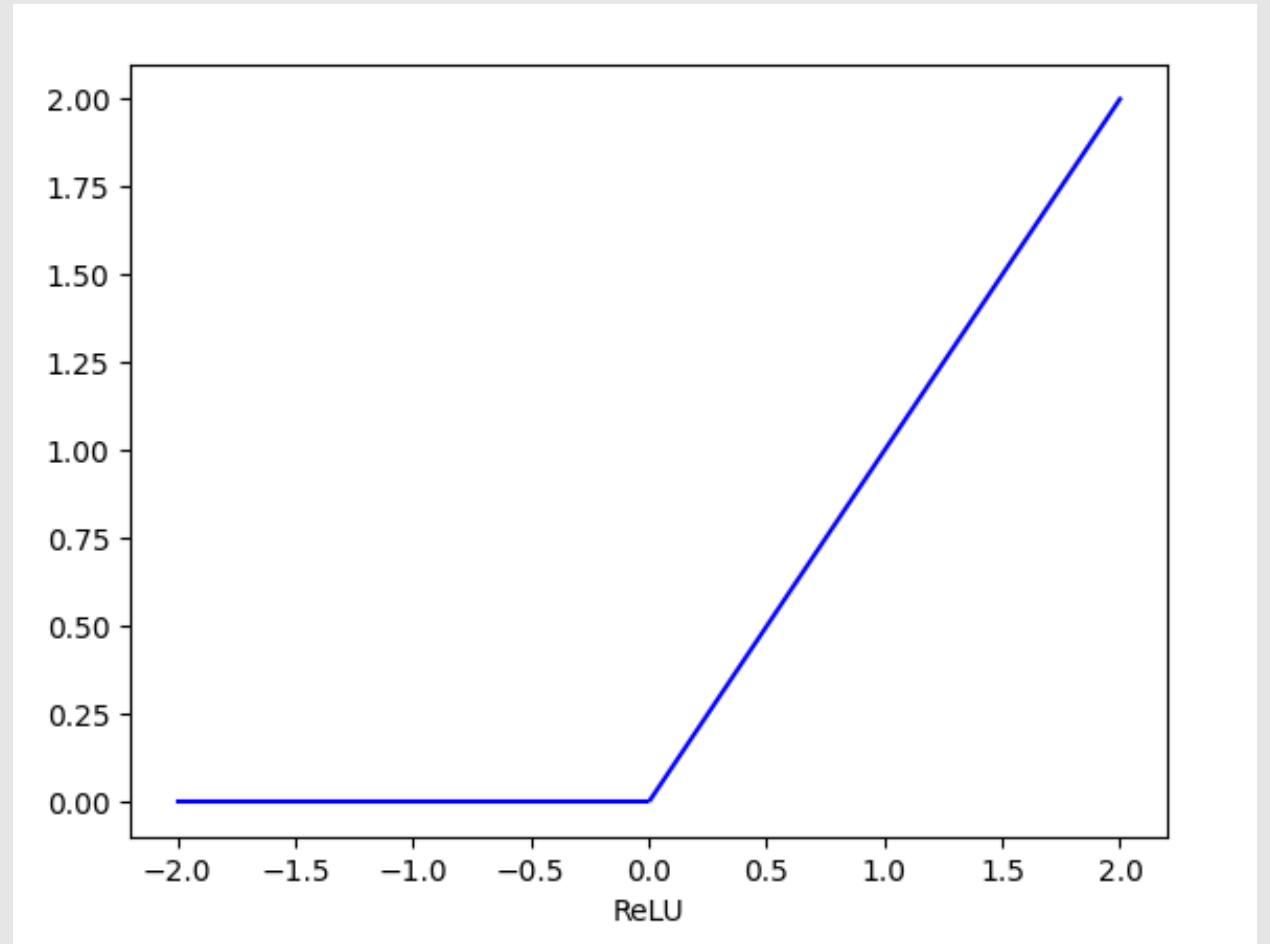
$$\text{ReLU}(x) = \max(0, x)$$

(Rectified Linear Unit)

利点

シグモイド関数では出力値が飽和してしまいう問題があった
入力値と出力値を同じ ($x > 0$)
にすることで解決

勾配の計算が容易



ReLU関数

交差検証 Cross validation

データをk個のスプリットに分割
(K-1)個のスプリットで訓練
残りの1つでテスト
全てのスプリットでテストを行い
平均する

汎化性能(未知のデータに対する
予測能力)を向上
結果を平均するため過学習のリス
クが少ない

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

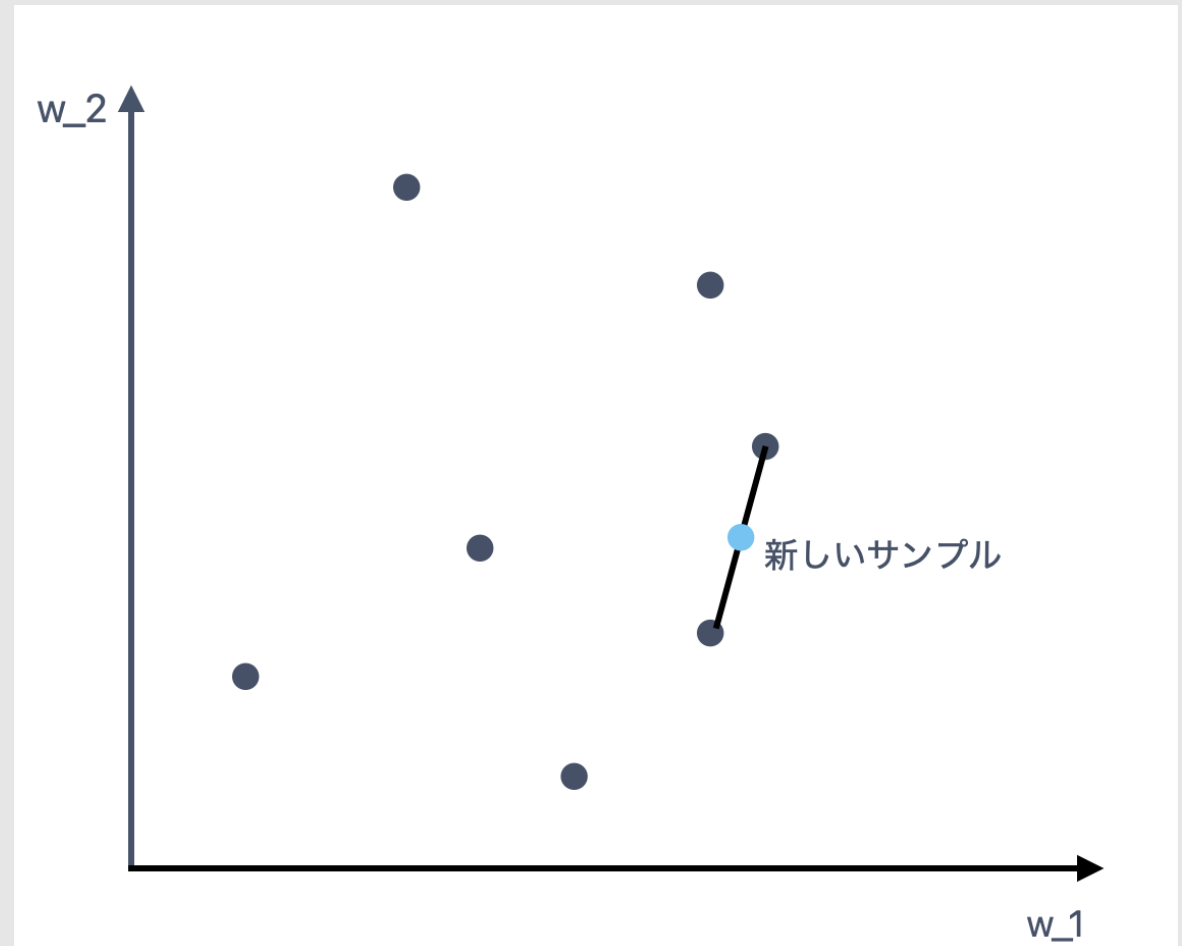
交差検証(k=5の例)

オーバーサンプリング SMOTE

SMOTE(Synthetic Minority Over-sampling Technique)

少数派クラスからランダムに
サンプルを選択
そのサンプルから距離の近い
(k 個)のサンプルを選び、
その直線上で新たなサンプル
を生成

右は $k=2$ の場合



SMOTE

使用したハイパーパラメータ

```
activation='relu', solver='adam',  
alpha=0.0001,  
batch_size='auto',  
learning_rate_init=0.01,  
early_stopping=True,  
beta_1=0.15, beta_2=0.990,  
epsilon=1e-10,  
n_iter_no_change=60, max_iter=600,  
hidden_layer_sizes=(30, 30),  
random_state=1,
```

欠損値を除いた場合の分類

