A-005

機械学習フレームワークへの識別学習機能追加のためのクラス設計
Class Design for Additional Function for Classification Learning into Machine Learning Framework

近藤 昌晴†
Masaharu Kondou

1. はじめに

現在広く使用されている scikit-learn では、ハイパーパラメータの直感的な間引き探索をする機能はない。

本発表では、回帰学習を持つ機械学習フレームワークに追加する形で、識別学習のハイパーパラメータ間引き探索を実施する機能と探索済みハイパーパラメータによる学習機能を設計・実装し、間引き探索によってパラメータ探索を高速化した結果を報告する。

2. 機能追加のベースとなる回帰学習器クラス

図 1 は、ハイパーパラメータの間引き探索処理と探索後のハイパーパラメータを持つ回帰学習を実施する HyperParamOptCV クラスを含むクラス図である。これにより、 (1) HyperParamOptCV クラス、(2) scikit-learn の回帰学習器 Mixin、そして (3) ベースとなる回帰学習器クラスを継承することでパラメータ探索と学習ができる。

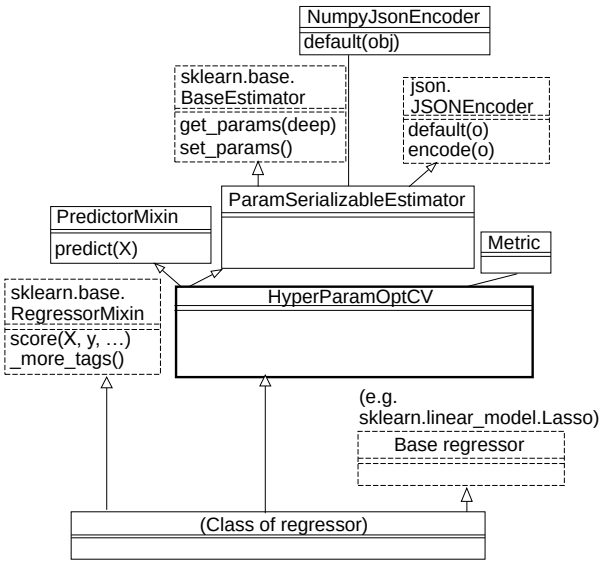

図 1　回帰学習器クラスのクラス図

但し、Gradient Boosting Machine（以下，GBM）系の機械学習フレームワークの回帰学習については、学習アルゴリズムで決定する内部パラメータを共通化させ、さらに決定された共通内部パラメータと探索対象のハイパーパラメータによる学習の手順を共通化させることができるため，共通パラメータ決定動作と共通学習手順をまとめるための GBM 回帰学習器クラスも定義する。図 2 に GBM 回帰学習のハイパーパラメータ探索・学習器クラスのクラス図を示す。

† 日立製作所研究開発グループ　Research & development group, Hitachi Ltd.

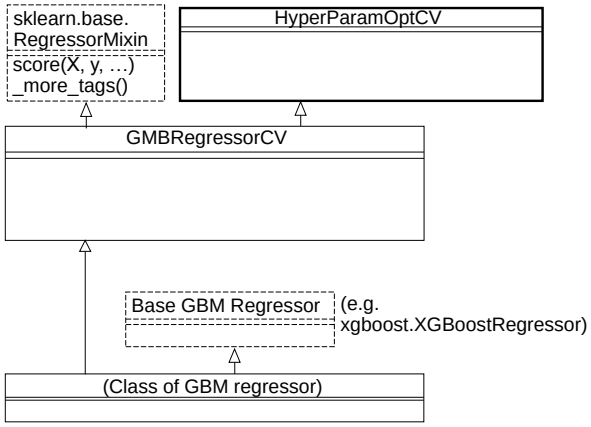

図 2　GBM 回帰学習器クラスのクラス図

3. 識別学習に対する機能追加の方針

識別学習に対する機能追加の基本設計も前述の回帰学習器同様のインターフェースを踏襲し、回帰学習と同様にパラメータ間引き処理・学習を利用できるようにする。しかし、回帰学習の機能をそのまま識別学習に適用できない課題が存在する。以下，課題とその対策方針を列挙する。

(1) 回帰学習用のメトリックは識別学習の学習精度評価に使用できない。対策として、識別学習用の評価メトリックをメトリック用クラスに追加する
(2) 識別学習における予測は、クラスの予測だけでなく、確率の予測と対数確率の予測の機能が必要である。対策として、Mixin クラスを新規に定義して継承する。
(3) ベースとする識別学習によっては，決定関数が scikit-learn のインターフェースとして用意されているものが存在する。対策として、これも Mixin クラスを新規に定義して継承する。
(4) 多クラス識別するベースの識別学習によっては、多クラス戦略（One-vs-All、多クラスアルゴリズム）を選択できるものが存在する。対策として、多クラス戦略用クラスを新規に定義する。

4. 設計・実装の結果

4.1 識別学習器クラスのクラス図

図 3 に、新たに設計・実装した識別学習器クラスのクラス図を示す。3 章の課題・対策方針に対応して修正または新規実装した部分をグレーで表示している。図 4 に、回帰学習と同様に共通パラメータ決定動作と共通学習手順をまとめるために設計・実装した GBM 識別学習器クラスのクラス図を示す。

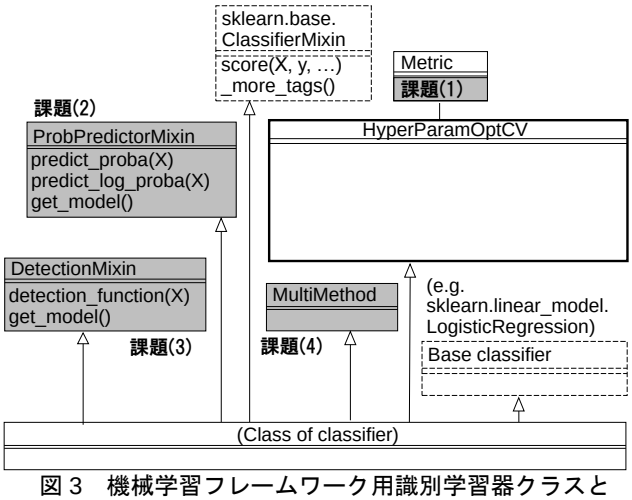

図 3　機械学習フレームワーク用識別学習器クラスと追加実装したサブクラスのクラス図

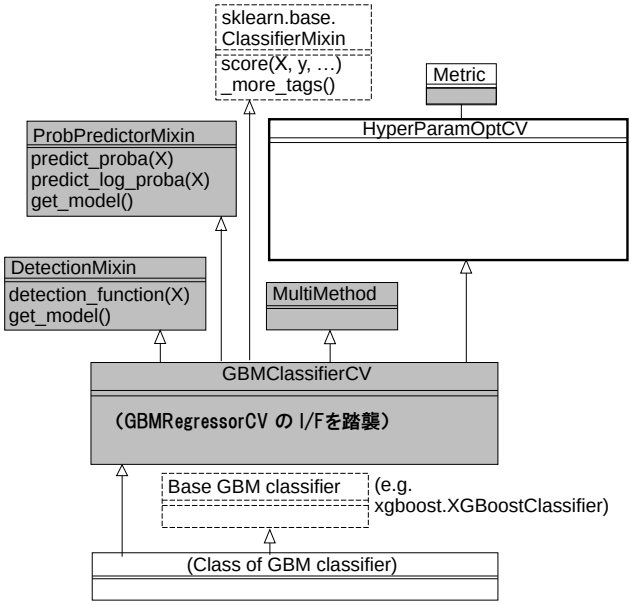

図 4　機械学習フレームワーク用 GBM 識別学習器クラスと追加実装したサブクラスのクラス図

4.2 パラメータの間引き処理の動作確認

表 1 に、実装済みのハイパーパラメータ間引き機能が探索するパイパーパラメータの範囲を示す。表 2 は、xgboost.XGBClassifier クラスと今回実装した GBMClassifierCV クラスを継承したクラスによる間引き探索と、ベースとなっている xgboost.XGBClassifier クラス単体の GridSearchCV による全探索について、探索するパラメータの最大組み合わせ数を合わせて比較した結果を掲載する。今回の実装によって、対象の機械学習フレームワークから識別学習のハイパーパラメータ探索を範囲指定なしで実施する場合、同じパラメータ範囲の単純な全探索と比較して探索数が 1/500 以下になる見込みを得た。

表 1　間引き探索対象となるパラメータ範囲

パラメータ名	探索範囲
max_depth	1, 3, 5, …, 29
min_child_weight	1, 3, 5, …, 9
gamma	0, 0.05, 0.1, … 0.95
subsample	0.2, 0.25, 0.3, … 0.95
colsample_bytree	0.2, 0.25, 0.3, … 0.95
reg_alpha	10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}

表 2　XGBoostClassifier パラメータの最大探索数

パラメータ名	最大探索数
実装インターフェース	3,072,000
GridSearchCV	5,203

5. おわりに

識別学習に対しても回帰学習と同様に機械学習フレームワークのパラメータ間引き探索を使用可能とする目的で、識別学習に対するハイパーパラメータ探索を実施する機能を設計・実装した。当該間引き探索機能を使用したときと、GridSearchCV によるパラメータ探索範囲の全組み合わせとの探索数を比較し、今回実装したインターフェースによってパラメータの最大探索数が 1/500 以下になる見込みを得た。

参考文献

[1] “scikit-learn,” http://scikit-learn.org/stable/ .
[2] “XGBoost, ” https://xgboost.ai/ .
[3] “LightGBM,” https://lightgbm.readthedocs.io/en/latest/ .
[4] “CatBoost,” https://catboost.ai/ .