MondrianForest テスト

MondrianForest のテストを行った。
http://archive.ics.uci.edu/ml/index.php というなんか色々データがあるサイトから落としてきた、以前学習用に遊んでいた adult というデータ(エクセルで修正)を使ってテストをしてみる。

データの内容は Age, Workclass, Fnlwgt, Education, Education_number, Marital_Status, Occupation, Relationship, Race, Sex, Capital_Gain, Capital_Loss, Hours_pre_week, Native_Country の 14 項目から、Imcome が>50K か、<=50K かを判別するというもの。
データ数は 32561 個で、テストデータは答え付きで 16281 個ある。
(計算時間なのでデータ数が多い方がいいと思ってこうしたのですが、処理が重すぎて死にました。勉強になりました。)

欠損値などあったので、seminar の P04S06_Scoring process-2 に従って get_dummies, RFE による特徴量選択などのデータの前処理をした。
特徴量は 15 個に絞り、残ったのは、Age, Fnlwgt, Education_number, Capital_Gain, Capital_loss, Hours_pre_week, Workclass_Self-emp-not-inc,…,など。

Partial_fit は pipeline できないため、事前に StanderdSchler による標準化を行い、「fit」など自体の処理を計測した。

まず、そもそもの性能を見るために、一度データ数に伴って各学習器の計算時間とスコアがどうなるかを調べる。ここでは、MondrianForest の Partial_fit は毎回リセットし処理を行っているので、蓄積されたデータ数がゼロの時から、データを入れるとどれくらい時間がかかるかを示している。

この時の accuracy score。

また、f1 score

最初の方でほぼ収束してしまっているので、軸を変えて。

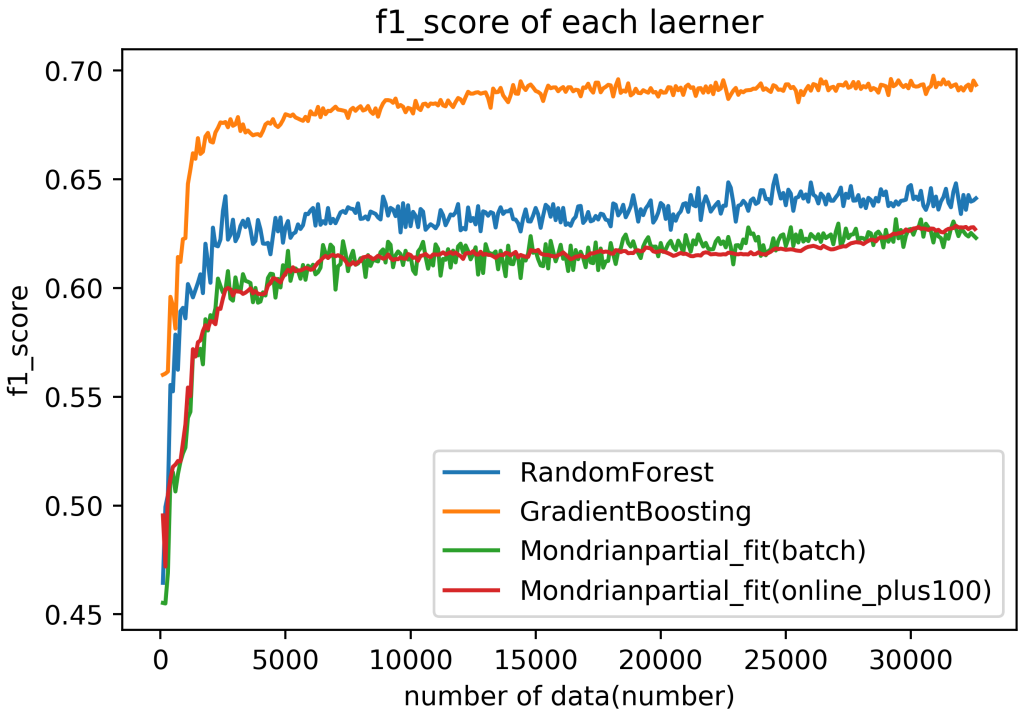
とりあえず、ここから読み取れることは、
・Mondrian の Partial_fit と fit はほぼ精度が変わらない(fit の方が若干高い)上に、partial_fit の方がだいぶ遅い。

・MondrianForest は RandomForest と比べるとそこまで精度に差はないが、GradientBoosting と比べると有意な差が出てしまっている。
GradientBoosting はα(学習率)など全てデフォルト値でやっているので、スピードより精度重視の設定になってはいるが。

次に、データが 1000 個づつ次々に到着して来たという状況を仮定する。
その場合の合計の計算時間を調べる。
ここでは、各学習器は(そこまでひどい状況はないと思うが)データが 1000 個追加するたびに再学習する。Mondrian の partial_fit は 1000 個来るたびに、追加学習する。

一応データが 100 個づつ来た場合。

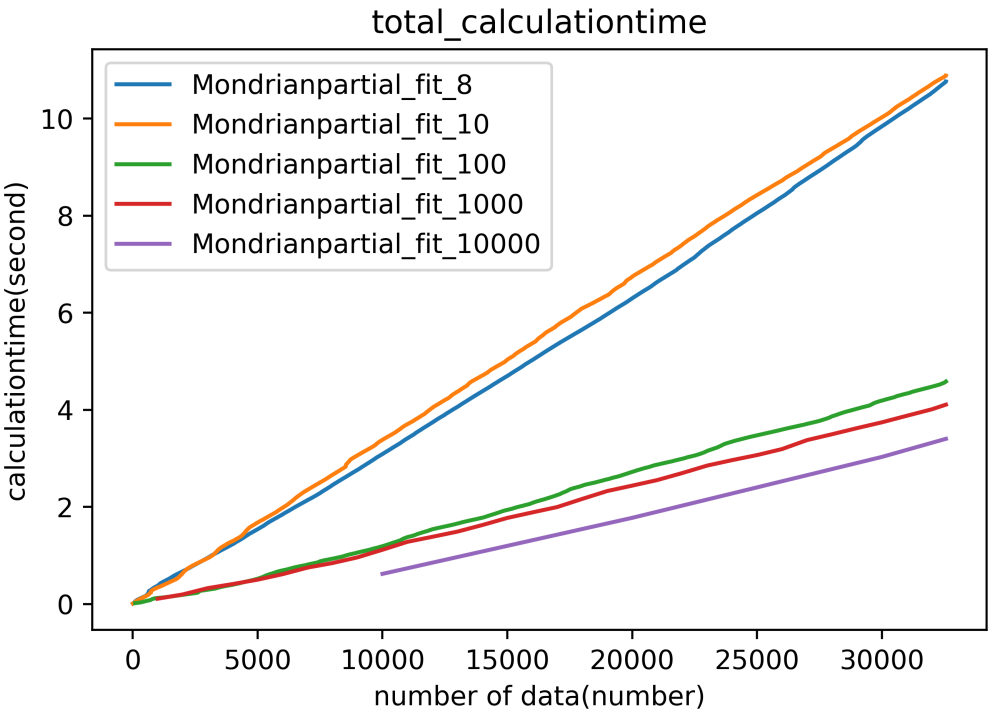
さすがに Mondrian が最も、そして大きく計算時間が早かった。
気になるのはこの時、精度が出ているかということ。
以下で、Mondrianpartial_fit(batch)は、最初に性能を見た時のようにデータ数がゼロから学習させたもの、Mondrianpartial_fit(online_plus100)は 100 個ずつデータを追加学習させたものである。



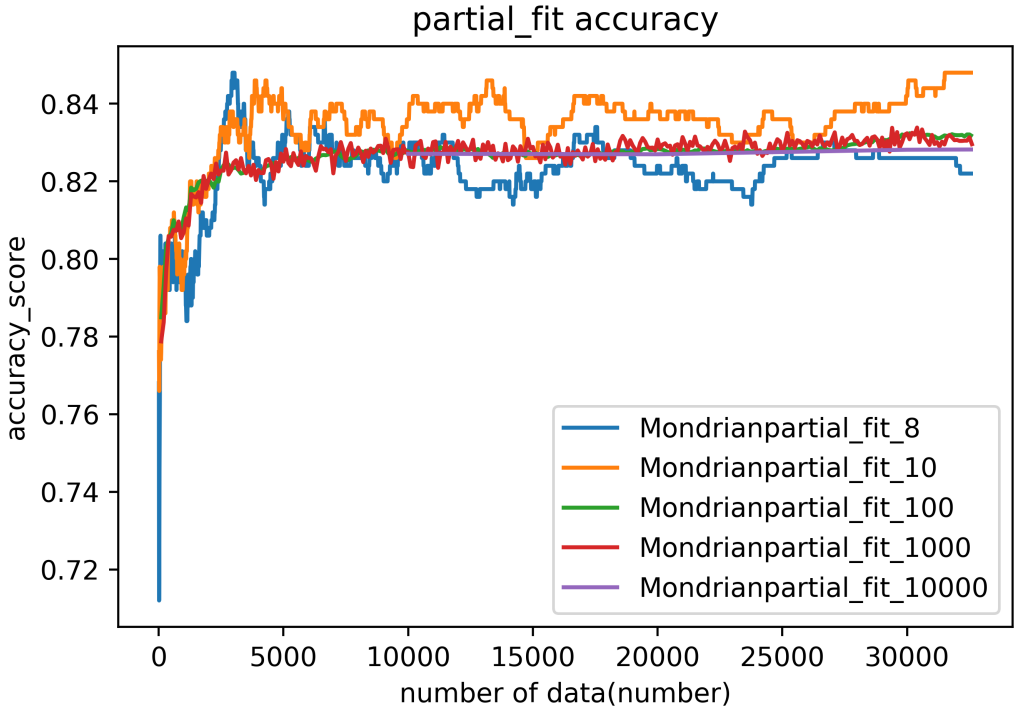
MondrianForest は一気に学習させた場合と追加学習させた場合でほぼ同じスコアであった。ちゃんと online 学習ができていると思われる。

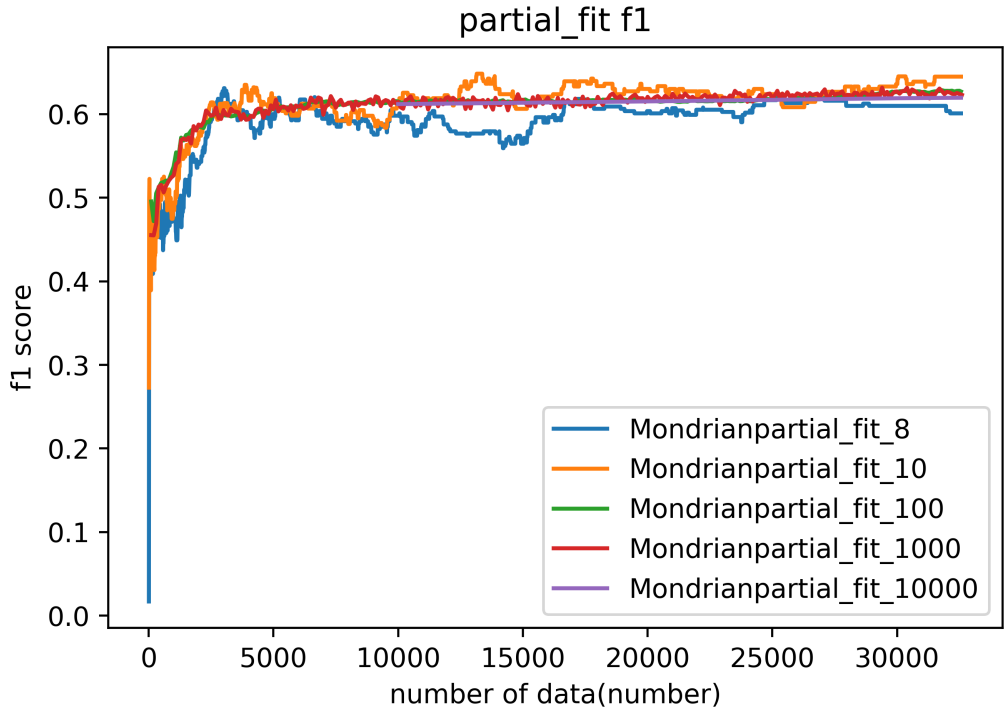
最後に、8 個ずつ、10 個ずつ、100 個ずつ、1000 個ずつ、10000 個ずつ足していった時の合計の計算時間を調べる。

8 個ずつなのは、一度に読み込ませるデータが 7 個以下だとエラーになるため 8 個ずつにした。

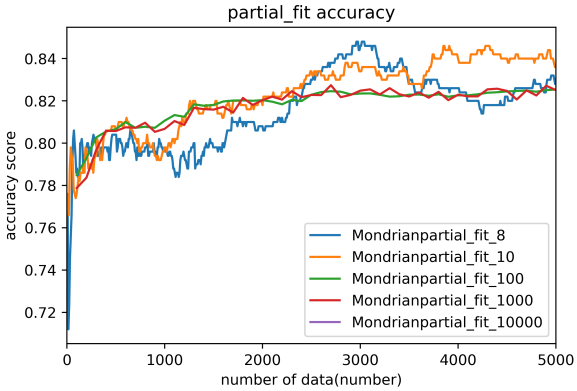


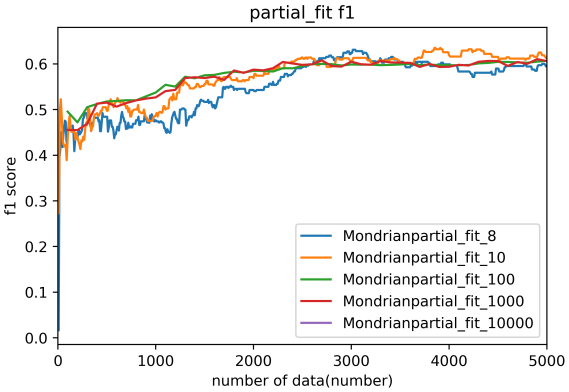
また、この時のスコアを示す。ただし、10 個ずつと 8 個ずつのものは、他のものはテストデータ 16281 個でスコアを計算しているが、計算が重すぎたので 500 個で行なっているため乱数性がある。


partial_fit accuracy
accuracy_score
0.84
0.82
0.80
0.78
0.76
0.74
0.72
0
5000
10000
15000
20000
25000
30000
number of data(number)
Mondrianpartial_fit_8
Mondrianpartial_fit_10
Mondrianpartial_fit_100
Mondrianpartial_fit_1000
Mondrianpartial_fit_10000


partial_fit f1
f1 score
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
5000
10000
15000
20000
25000
30000
number of data(number)
Mondrianpartial_fit_8
Mondrianpartial_fit_10
Mondrianpartial_fit_100
Mondrianpartial_fit_1000
Mondrianpartial_fit_10000

軸を変えて





一気に多くのデータを読み込ませた方が時間が少ないことがわかった。
また、先ほども同じようなものを示したが、このデータではデータを一度に取り込む量とスコアに相関性はなかった。

まとめ。

・MondrianForest の partial_fit や fit 自体の計算時間は早くはなく、fit の方が早かった。

・MondrianForest は精度に少し懸念がある。

・MondrianForset で一気に学習させた場合(batch)と追加学習の場合(online)では精度に差はみられなかった。

・MondrianForest では一度に読み込ませるデータ数で精度の変化がみられず。また、一度にたくさんのデータを読み込ませた方が計算が早かった。