Boruta

明治大学 理工学部 応用化学科
データ化学工学研究室　金子 弘昌

Boruta とは？

✓ランダムフォレスト (Random Forest, RF) の変数重要度に基づく変数選択手法

- RF についてはこちら: https://datachemeng.com/randomforest/

✓あえて目的変数と関係のない説明変数を追加して RF を行い、その変数重要度と、オリジナルの説明変数における変数重要度とを比較することで、選択する説明変数を検討

Boruta に着目した理由

- ✓Stepwise や GAPLS などの多くの変数選択手法は、クロスバリデーション後の r^2 などの何らかの統計量を最適化するように変数を選択する
  - Stepwise: https://datachemeng.com/stepwise/
  - GAPLS: https://datachemeng.com/gaplsgasvr/
- ✓変数選択前と比べて統計量は改善されるが、外部データに対する推定性能は考慮されてなく、オーバーフィットする危険がある

- ✓Boruta では統計量の最適化をしていないため、オーバーフィッティングの影響を軽減できるかも！？

Boruta のアルゴリズム 1/3

1. 説明変数のデータセットである m × n の行列 (m はサンプル数、n は説明変数の数) をコピーする

2. 1. でコピーした行列において、変数ごとにサンプルの値をランダムに並び替える
   - ✓ここで準備した変数をランダム説明変数と呼ぶことにします。変数ごとに値をランダムに並び替えているため、目的変数と関係はありません
   - ✓ランダム説明変数のデータセットは m × n の行列です

3. オリジナルの説明変数のデータセットと、ランダム説明変数のデータセットとを一緒にして、目的変数との間で RF を実行し、変数重要度を計算する

Boruta のアルゴリズム 2/3

4. ランダム説明変数における変数重要度 (n 個) の、
   p パーセンタイルを基準値とする
   - ✓ランダム説明変数は目的変数と関係ありませんが、何らかの値が変数重要度として割り当てられます。重要でない説明変数によって変数重要度の基準値を設けます
   - ✓オリジナルの Boruta では p = 100、つまり最大値です

5. オリジナルの説明変数において、変数重要度が 4. の基準値を越えた変数を hit とする
   - ✓目的変数と関係のないランダム説明変数の変数重要度の基準値は越えていてね、ということです

Boruta のアルゴリズム 3/3

6. 2.-5. を繰り返す中で、両側検定でオリジナルの説明変数がランダム説明変数と比較して重要かどうか検討する
   - ✓hit したか否かなので、二項分布です
   - ✓オリジナルでは有意水準 α = 0.05 です
   - ✓2.-5. の繰り返しの中でも、ランダム説明変数と比較して重要でないと判断されたオリジナルの説明変数は削除されます

Python で Boruta を実行するには？

✓boruta_py

- https://github.com/scikit-learn-contrib/boruta_py
- https://pypi.org/project/Boruta/

Boruta のパラメータ

- ✓RF の設定として、用いる説明変数の割合を、0.1, 0.2, …, 0.9 と振って、Out-Of-Bag で最適化

- ✓p = 100 とすると、説明変数が削除されすぎて、モデルの推定性能が低下することがある
- ✓特にサンプルが少ないときなど、ランダムに並び替えたといってもたまたま目的変数と関係性が出てきてしまう変数もあることを想定して、変数をランダムに並び替えて目的変数と相関係数を計算することを 10000 回くらい行い、その相関係数の絶対値の最大値を r_{ccmax} としたとき、

  $p = 100 \times (1 - r_{ccmax})$

  とするとよさそう

解析してみました 比較手法

- ✓Boruta を用いた変数選択前後で、以下の回帰分析手法によりモデル構築した結果を比較

- ✓Partial Least Squares (PLS)
- ✓Ridge Regression (RR)
- ✓Least Absolute Shrinkage and Selection Operator (LASSO)
- ✓Elastic Net (EN)
- ✓Support Vector Regression (SVR) [ガウシアンカーネル]
- ✓Decision Tree (DT)
- ✓Random Forest (RF)
- ✓Gaussian Process regression (GP)
- ✓Light GBM (LGB)
- ✓XGBoost (XGB)
- ✓Gradient Boosting Decision Tree (GBDT)
  - 各手法の詳細はこちら: https://datachemeng.com/summarydataanalysis/

解析してみました 記述子

✓記述子は RDKit で計算

- https://www.rdkit.org/docs/api-docs.html

✓トレーニングデータとテストデータとに分割したあとに標準偏差が 0 の記述子は事前に削除

解析してみました 沸点のデータセット

✓沸点のデータセット

- Lowell H. Hall, C. T. Story,
  J. Chem. Inf. Comput. Sci. 1996, 36, 1004 - 1014

✓沸点が測定された 294 個の化合物

✓トレーニングデータ: 220 化合物

✓テストデータ: 74 化合物

✓記述子の数: 144

✓Boruta によって選択された記述子の数: 81 (56 %)

解析してみました 沸点のデータセット 推定結果

✓テストデータの r^2, RMSE

	変数選択前		Borutaで選択後	
	r^2	RMSE	r^2	RMSE
PLS	0.817	34.5	0.824	33.9
RR	0.781	37.8	0.841	32.2
LASSO	0.770	38.8	0.832	33.1
EN	0.775	38.3	0.841	32.2
SVR	0.846	31.8	0.886	27.3
DT	0.807	35.5	0.813	35
RF	0.838	32.5	0.843	32
GP	0.851	31.2	0.927	21.8
LGB	0.813	34.9	0.807	35.5
XGB	0.848	31.5	0.848	31.5
GBDT	0.861	30.1	0.863	29.9

解析してみました 環境毒性のデータセット

✓環境毒性のデータセット
- http://www.cadaster.eu/node/65.html

✓環境毒性が測定された 1213 個の化合物

✓目的変数である $pIGC_{50}$ とは、ある時間に Tetrahymena pyriformis の増殖の 50 % を阻害する化合物の濃度を IGC_{50} [μM]としたときの $-\log(IGC_{50})$

✓トレーニングデータ: 910 化合物

✓テストデータ: 303 化合物

✓記述子の数: 164

✓Boruta によって選択された記述子の数: 78 (48 %)

解析してみました 環境毒性のデータセット 推定結果

✓テストデータの r^2, RMSE

	変数選択前		Borutaで選択後	
	r^2	RMSE	r^2	RMSE
PLS	0.768	0.509	0.734	0.545
RR	0.777	0.499	0.762	0.515
LASSO	0.783	0.492	0.744	0.535
EN	0.782	0.494	0.757	0.521
SVR	0.814	0.456	0.820	0.449
DT	0.652	0.624	0.641	0.633
RF	0.799	0.474	0.784	0.491
GP	0.807	0.465	0.814	0.456
LGB	0.825	0.442	0.800	0.473
XGB	0.818	0.451	0.792	0.483
GBDT	0.807	0.464	0.786	0.489

解析してみました 薬理活性のデータセット

✓薬理活性のデータセット

- Jeffrey J. Sutherland, Lee A. O'Brien, Donald F. Weaver, J. Med. Chem., 2004, 47(22), 5541-5554

✓アンジオテンシン変換酵素阻害薬 (高血圧 (血圧上昇) および うっ血性心不全の治療に使用される医薬品) として薬理活性が 測定された 114 個の化合物

✓目的変数である pIC_{50} とは、標的のものの 50 % を阻害する 化合物の濃度を IC_{50} [μM]としたときの $-\log(IC_{50})$

✓トレーニングデータ: 86 化合物

✓テストデータ: 28 化合物

✓記述子の数: 146

✓Boruta によって選択された記述子の数: 70 (48 %)

解析してみました 薬理活性のデータセット 推定結果

✓テストデータの r^2, RMSE

	変数選択前		Borutaで選択後	
	r^2	RMSE	r^2	RMSE
PLS	0.689	1.323	0.840	0.949
RR	0.872	0.850	0.845	0.935
LASSO	0.793	1.081	0.845	0.933
EN	0.841	0.947	0.844	0.937
SVR	0.777	1.122	0.794	1.076
DT	0.748	1.193	0.748	1.193
RF	0.863	0.879	0.862	0.881
GP	0.879	0.826	0.858	0.893
LGB	0.861	0.885	0.862	0.882
XGB	0.854	0.906	0.847	0.928
GBDT	0.850	0.919	0.838	0.956

解析してみました 融点のデータセット

✓融点のデータセット
- Karthikeyan, M., Glen, R. C., Bender, A.,
  J. Chem. Inf. Model., 45(3), 581–590. 2005

✓融点が測定された 4333 個の化合物

✓トレーニングデータ: 1,000 化合物

✓テストデータ: 3,333 化合物

✓記述子の数: 187

✓Boruta によって選択された記述子の数: 70 (37 %)

解析してみました 融点のデータセット 推定結果

✓テストデータの r^2, RMSE

	変数選択前		Borutaで選択後	
	r^2	RMSE	r^2	RMSE
PLS	0.394	49.3	0.338	51.5
RR	0.414	48.5	0.35	51.0
LASSO	0.387	49.6	0.349	51.1
EN	0.390	49.4	0.338	51.5
SVR	0.514	44.1	0.490	45.2
DT	0.221	55.9	0.225	55.7
RF	0.457	46.6	0.453	46.8
GP	0.510	44.3	0.501	44.7
LGB	0.471	46.0	0.460	46.5
XGB	0.459	46.6	0.453	46.8
GBDT	0.461	46.5	0.452	46.9

解析してみました 水溶解度のデータセット

- ✓水溶解度のデータセット
  - Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J., J. Chem. Inf. Comput. Sci. 2004, 44, 266–275.
- ✓水溶解度が測定された 1290 個の化合物
- ✓目的変数である logS とは、水への溶解度を S [mol/L] としたときの log(S)

- ✓トレーニングデータ: 968 化合物
- ✓テストデータ: 322 化合物

- ✓記述子の数: 186
- ✓Boruta によって選択された記述子の数: 93 (50 %)

解析してみました 水溶解度のデータセット 推定結果

✓テストデータの r^2, RMSE

	変数選択前		Borutaで選択後	
	r^2	RMSE	r^2	RMSE
PLS	0.896	0.694	0.880	0.745
RR	0.901	0.679	0.890	0.713
LASSO	0.899	0.685	0.888	0.719
EN	0.901	0.677	0.890	0.715
SVR	0.923	0.599	0.916	0.623
DT	0.876	0.757	0.883	0.736
RF	0.925	0.588	0.924	0.592
GP	0.924	0.595	0.919	0.613
LGB	0.928	0.579	0.923	0.599
XGB	0.925	0.588	0.925	0.591
GBDT	0.928	0.579	0.926	0.587

参考文献

- ✓ Kursa M., Rudnicki W., "Feature Selection with the Boruta Package“, Journal of Statistical Software, Vol. 36, Issue 11, Sep 2010

- ✓ http://danielhomola.com/2015/05/08/borutapy-an-all-relevant-feature-selection-method/