Python による

TCGA BRCA に基づく Basal Subtype のバイオマーカー構築

Via PCA + Random Forest

2018 年 9 月 ２１日
臨床薬理部
前田公介

公開データとして，TCGA にて公開されているデータを用いた。解析ソフトは、Python を用いた。

機械学習の方法としては、主成分分析、ランダムフォレスト法を採用した。

TCGA Breast Cancer Data の分類問題（PCA + RandomForest Model）

解析対象データ：TCGA 乳癌データ（１１０４例）

Basal Subtype をクラス 0 とし、非 Basal Subtype をクラス 1 とした。2 クラスの分類問題である。

説明変数としては、代表的な癌幹細胞マーカー及び EMT マーカーを採用し、これらの変数に基づいて、Basal Subtype のクラスの有無を分類するという機械学習の解析を実施した。

解析に採用した癌幹細胞及び EMT マーカーの一覧

CDH2 CDH3 CDK9 CDKN1A IGF2BP1 IGF2BP3 IGF2BP2 CTCF COL1A1 COL3A1 CLDN7 CSF2 CTGF CTNNB1 CTSL CYP24A1 GADD45A JAG1 EGFR ELAVL1 ERBB2 ERBB3 ERBB4 ESR1 EZH2 ALDH1A1 ALDH1B1 FAP ALDH1A3 DKK1 FOXM1 FLNA FN1 BRD4 LDLRAP1 GATA3 GLI1 GLI2 DLL1 CXCL1 HAS2 APC SNAI3 IDI1 IGF1R IGF2 IGFBP3 IGFBP5 IGFBP6 IL6 CXCL8 INSIG1 ITGA6 AR ITGA3 ITGAV ITGB4 ITGB5 KRT5 KRT8 KRT14 KRT18 KRT19 LIN28B LCN2 LRP6 EPCAM MCM2 MCM3 MCM4 MCM5 MCM6 MCM7 MET MMP2 MMP9 MMP13 MYC NOTCH1 NSDHL LEF1 PDGFA PDGFB PECAM1 PGR PIGR PLEC POU1F1 POU5F1 FBXW7 PDGFC PTEN PTPN1 PTPN3 ACTA1 ACTA2 RB1 CCND1 ACTB CCL2 CXCL12 SHH SNAI2 SNAI1 SOX2 SPARC BRCA1 BRAF BRCA2 TCF4 TCF7L2 ZEB1 TGFB1 TGFB1I1 TGFB2 TGFB3 TGFBI THBS1 TIMP1 TIMP2 TJP1 TP53 TP53BP1 TP73 TWIST1 VCL VDR VEGFA VEGFB VEGFC EZR VIM WNT2B WNT9A CXCR4 NANOG PDGFD HMGA2 AXIN1 AXIN2 LGR5 TP63 PROM1 ALDH1A2 CLDN12 CLDN2 CLDN1 DCLK1 AURKB CYTH3 TJP2 CD44 GIT2 CDK1 ZEB2 CDH1

（以上）

作成した Python 解析プログラムファイルの一覧

ファイル名	内容
0_Analysis_Data_Make.py	解析データの構築
2_PCA Transformed Plot.py	主成分分析後の PC1，PC2 のプロット
3_PCA cumulative explained variance.py	各主成分の寄与率
4_Random Forest after PCA transformation_Decision Region.py	PCA+RandomForest の決定領域の図示
4_Random Forest after PCA transformation_Matrix.py	PCA+RF のモデルの検証
5.1_Model Evaluation_learning Curve.py	学習曲線
5.2_Model Evaluation_Validation Curve_n_estimators.py	RF の n_estimators の検証曲線
5.3_Model Evaluation_Validation Curve_max_depth.py	RF の max_depth の検証曲線
5.3_Model Evaluation_Validation Curve_n_componetns.py	PCA の n_components の検証曲線
5.4_Model Evaluation Confusion Matrix.py	正答率の図示
6.1_ROC Curve.py	ROC 曲線の図示

（1）解析対象データの構築

データの構造は以下の通りである。各患者に対して、列名に遺伝子名として、行名に患者のＩＤを付している。また、クラスとして Basal Type と non-Basal Type の 2 クラスを作成している。

解析データの構成

Index	CDH2	CDH3	CDK9	CDKN1A	IGF2BP1	IGF2BP3	IGF2BP2	CTCF	COL1A1	COL3A1
TCGA-3C…	2.13556	3.18778	43.9001	74.3301	0.136807	0.284249	0.687948	66.3435	937.251	576.493
TCGA-3C…	1.9746	16.1965	55.2016	57.095	0.625794	0.388671	6.60516	40.4985	2522.13	1093.46
TCGA-3C…	2.3319	6.33212	60.384	164.839	0.236543	0.364225	0.948312	36.012	2418.75	1008.4
TCGA-3C…	39.528	19.041	36.1484	128.038	0.129748	0.176953	0.459879	30.868	5832	2847.75
TCGA-4H…	3.33202	13.2628	56.2576	67.3961	0.114239	0.246822	2.41743	41.0679	6560.29	4110.93
TCGA-5T…	0.336078	9.24131	36.816	42.5828	0.345616	0.144761	0.522076	51.6864	99.9113	42.7858
TCGA-A1…	1.33134	50.582	32.1931	63.2835	0.1	0.253487	0.956148	47.7076	231.186	623.928
TCGA-A1…	6.02677	51.3242	49.5213	123.421	0.167716	0.344199	1.73958	48.8765	1672.96	1255.47
TCGA-A1…	1.41967	31.531	40.2609	66.7423	0.1	0.168996	1.04205	44.9136	3309.97	2990.99
TCGA-A1…	0.757811	23.0206	43.2913	32.6328	0.123368	0.243414	0.513531	42.323	1225.01	1328.8
TCGA-A1…	2.03281	11.4191	44.7537	281.827	0.126404	0.325455	1.0284	47.4253	2513.04	1700.38
TCGA-A1…	10.014	21.1291	36.167	69.7297	0.360965	0.256828	4.18489	37.6888	7181.49	5873.5
TCGA-A1…	3.032	11.9711	46.7879	99.8422	0.129896	0.33616	0.603862	48.3466	7181.49	5873.5
TCGA-A1…	3.62585	19.6621	24.7726	77.3528	0.114853	0.248054	0.91218	59.0464	1056.64	1037.65
TCGA-A1…	39.528	17.1506	27.8917	219.01	0.213986	0.261773	4.51442	53.6617	380.573	305.182
TCGA-A1…	20.1874	7.4334	48.8421	144.396	0.172589	0.402999	2.69354	52.1068	3961.87	3264.71
TCGA-A1…	1.4461	13.774	21.7868	8.9607	0.106827	0.196336	0.14406	66.882	247.341	229.954
TCGA-A1…	39.528	42.6364	25.6897	70.5182	0.158937	7.91696	16.1947	35.4494	759.507	629.534

患者のクラス値

Index	PAM50lite
TCGA-A1-A0SE	Non-basal
TCGA-A1-A0SF	Non-basal
TCGA-A1-A0SG	Non-basal
TCGA-A1-A0SH	Non-basal
TCGA-A1-A0SI	Non-basal
TCGA-A1-A0SJ	Non-basal
TCGA-A1-A0SK	Basal
TCGA-A1-A0SN	Non-basal
TCGA-A1-A0SO	Basal
TCGA-A1-A0SP	Basal
TCGA-A1-A0SQ	Non-basal
TCGA-A2-A04N	Non-basal
TCGA-A2-A04P	Basal
TCGA-A2-A04Q	Basal

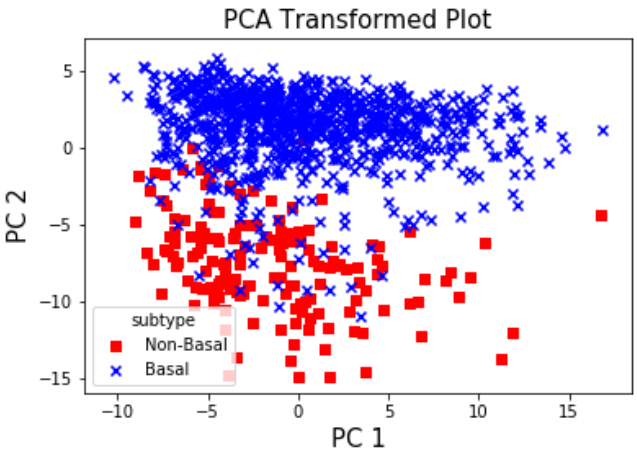
（２）遺伝子発現データのプロット

省略

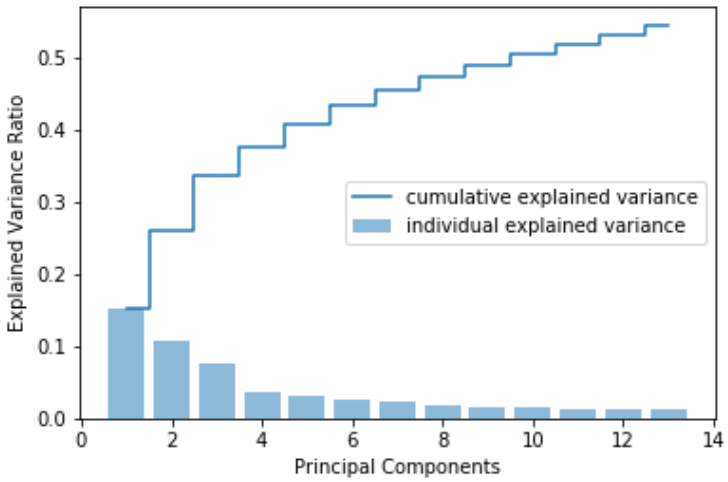
（３）PCA 変換後のデータ

オリジナル・データに対して主成分分析解析を実施して、第１主成分及び第２主成分をＸ軸、Ｙ軸として散布図を作成したところ、以下のとおりであった。

Basal Subtype と non-Basal Subtype が、適切に分離されていることが推測される。主成分分析でデータ変換をしてから、機械学習を実施することが、望ましいと判断される。



第１主成分と第２主成分の両方で、全体の分散のうち、30%程度を説明できることが示された。



（４）PCA 変換後データに対する Random Forest モデルの適用

オリジナル・データのうち，８０％を学習データとして用いて、残りの２０％を検証データとして採用した。このデータの抽出は、無作為に行った。

Random Forest モデルの適用の前処理としては、以下の手順に従った。

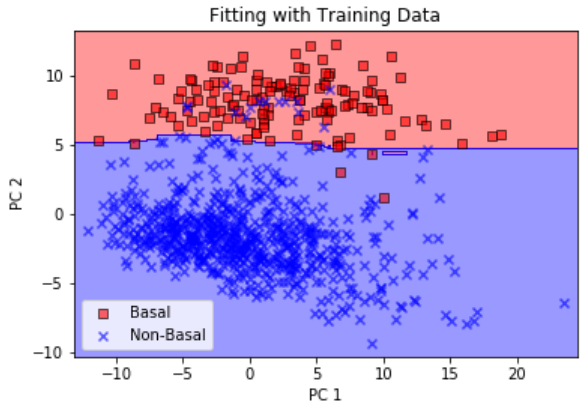
①対数化
②正規化及び標準化
③PCA 変換

モデルパラメータは以下の通りと設定した。

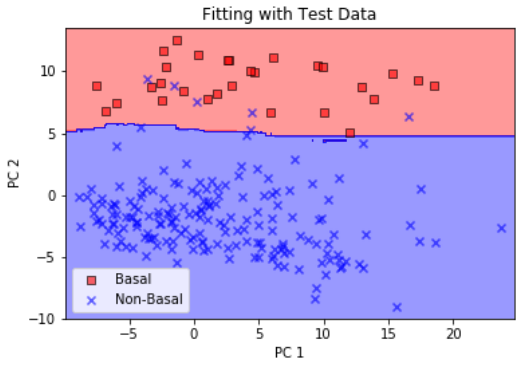
パラメーター	値	
n_component	2	
n_estimators	20	
max_depth	5	

学習データに基づいてモデルを構築し，モデルに基づいて，分割曲面を作成したのが以下の図である。青色と赤色の分割は、モデル値に基づいている。実際のデータを重ねてプロットしたところ、学習データでは、適切に分類されていることが視察的に示された。

学習データに基づいて構築されたモデルの予測値と学習データとを重ね図であるため、当然、検証データよりもよくフィットしている。



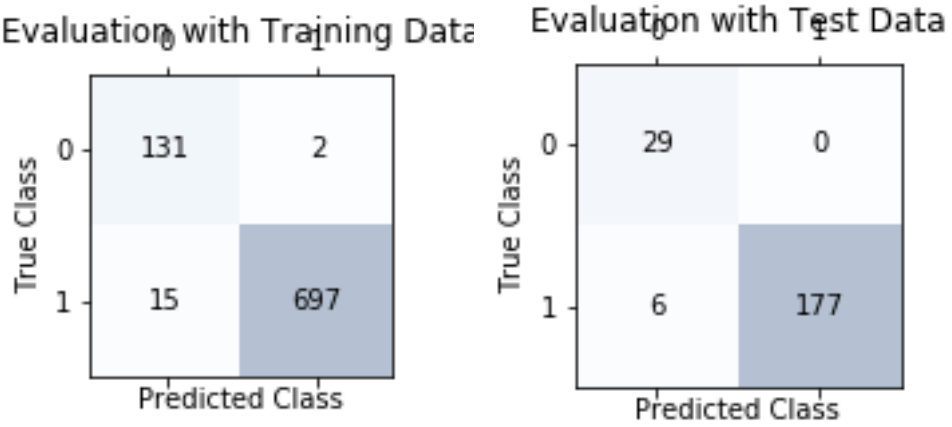
次に，学習データに基づいて構築されたモデルにより、検証データでの妥当性を確認した。学習データに比べれば正答率は低下しているものの、検証データにおいても、ほぼ適切に分類できていることが示された。



第 2 主成分が 5 以上のときには、Basal Type と判断していいことが示された。

次に、学習データ及び検証用データでの正答率を算出した。学習データにおけるモデルの正答率は、９８％であり、検証用データにおけるモデルの正答率は、９７．２%であった。

学習データに基づいて構築したモデルによる予測値との正答率であるため、学習データにおける正答率の方が、検証用データにおける正答率よりも良いのは当然である。

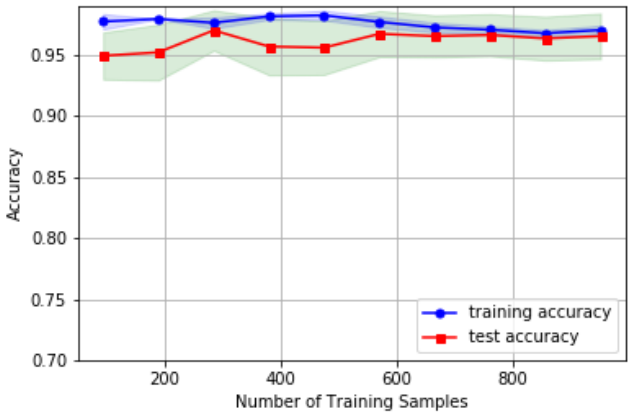


Accuracy with Training Data: 0.980

Acccuracy with Test Data: 0.972

（５）PCA 変換後データに対する Random Forest モデルの学習曲線

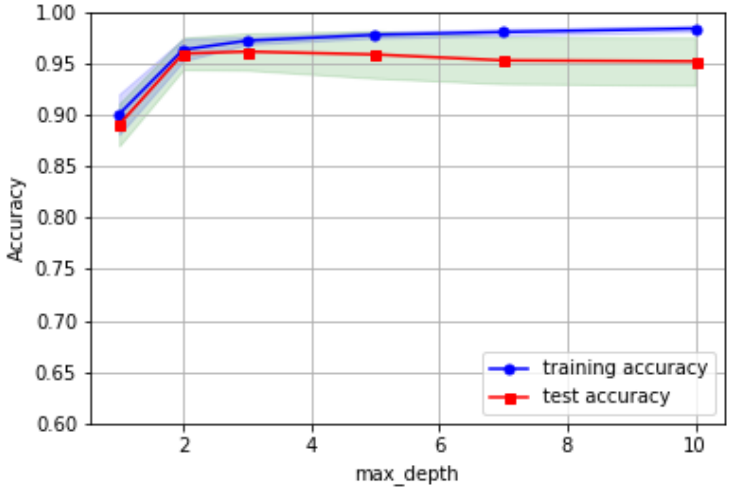
次に，学習データのサンプル数を任意に変更した場合における正答率の推移を調べた。全体が１０１４例であるが、学習データの例数は低くても、高い正答率を確保できることが示された。



（６）PCA 変換後データに対する Random Forest モデルの検証曲線

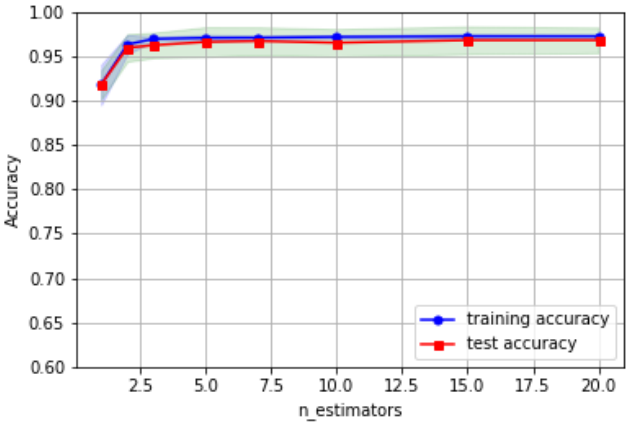
① Random Forest の max_depth についてのチューニング

Max_depth は、3 の時点で正答率は、飽和していることが示された。
最終モデルとしては、max_depth は 3 とする。



② Random Forest の n_estimators についてのチューニング

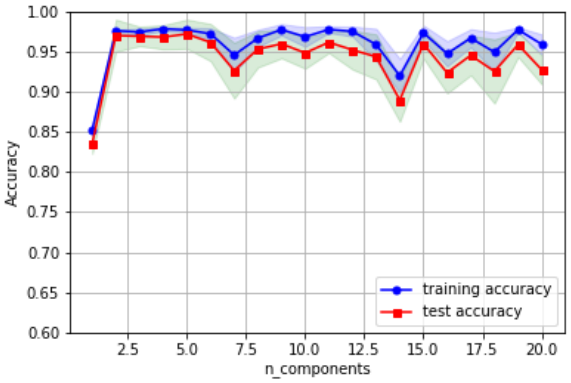
Max_depth は、５の時点で正答率は、飽和していることが示された。
最終モデルとしては、n_estimators は５とする。



② ＰＣＡのについてのチューニング

N_components は、2 の時点で正答率は、飽和していることが示された。
最終モデルとしては、n_component は 2 とする。

ＰＣＡプロットで、Ｘ軸を PC1 とし、Ｙ軸を PC2 としてクラス分類のプロットをした際にも、Basal Type と non-Basal Type とで適切に分類されていたことと合致する。



以上から、最終モデルのモデルパラメータとして以下の通りとなった。

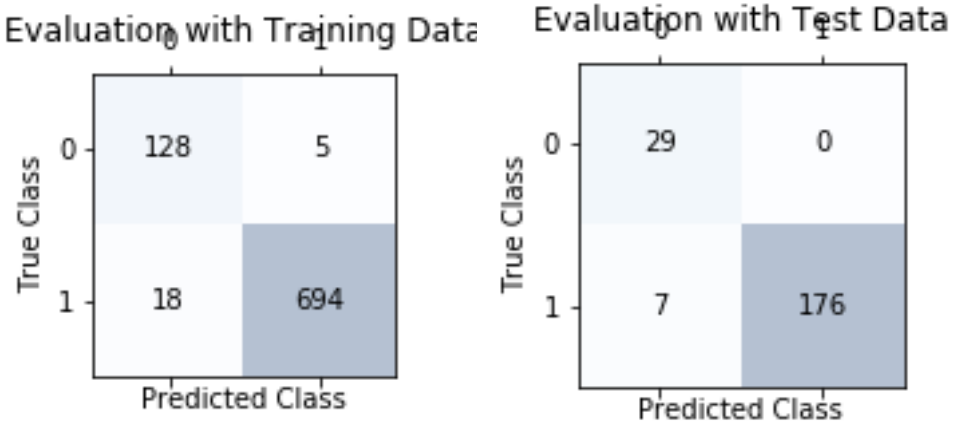
最終モデルパラメータ

パラメーター	値	
n_component	2	
n_estimators	5	
max_depth	3	

（７）最終モデルに基づく正答率の算出

（６）にて記載した最終モデルに基づいて解析を行ったところ、学習データにおける正答率は、９７．３％であり、検証データにおける正答率は、９６．７％であった。

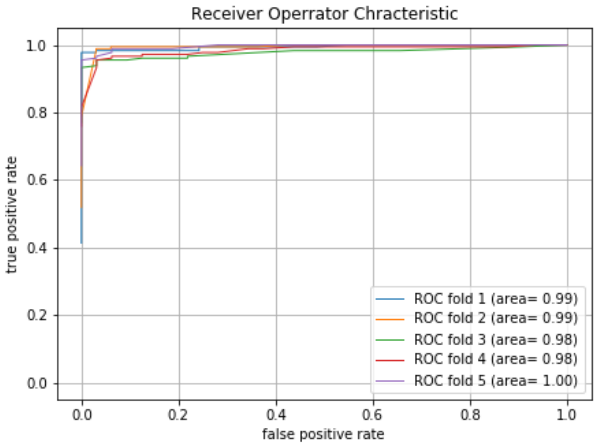
モデルパラメータの数を増やせば増やすほど、過学習の問題が生じるため、この最終モデルが適切であると判断される。



Accuracy with Training Data: 0.973
Acccuracy with Test Data: 0.967

（８）PCA 変換後データに対する最終モデルの ROC 曲線

全体の１１０４例のうち、５分割法により、５分の４を学習データとして用いて、残りの５分の１を検証用データとして真陽性率、偽陽性率をそれぞれ求めた（5 分割交差検証法)。５パターンを算出し、それぞれを重ねてプロットした図が以下の通りである。５パターンの AUC の平均は、0.99 であり、かなり予測が良いことが示された。



（９）結語

乳癌の Basal Subtype として、癌幹細胞マーカー及びＥＭＴマーカーが重要であることが知られている。

最終モデルの検証の結果，学習データの正答率が９７．３％、検証データの正答率が９６．７％であった。

ＲＯＣ曲線の AUC は、０．９９であったことから、このバイオマーカーの樹立にほぼ成功したものと考えられる。

乳癌の Basal Subtype では、癌幹細胞遺伝子や EMT 関連遺伝子に特徴があることが示された。

また、膨大な遺伝子数からバイオマーカーを樹立する際には、主成分分析を施し、その結果、得られた第１主成分と第２主成分を説明変数として、ランダムフォレスト法を実施するという手法は、汎用的に有効な方法であることが示唆された。

以上