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）

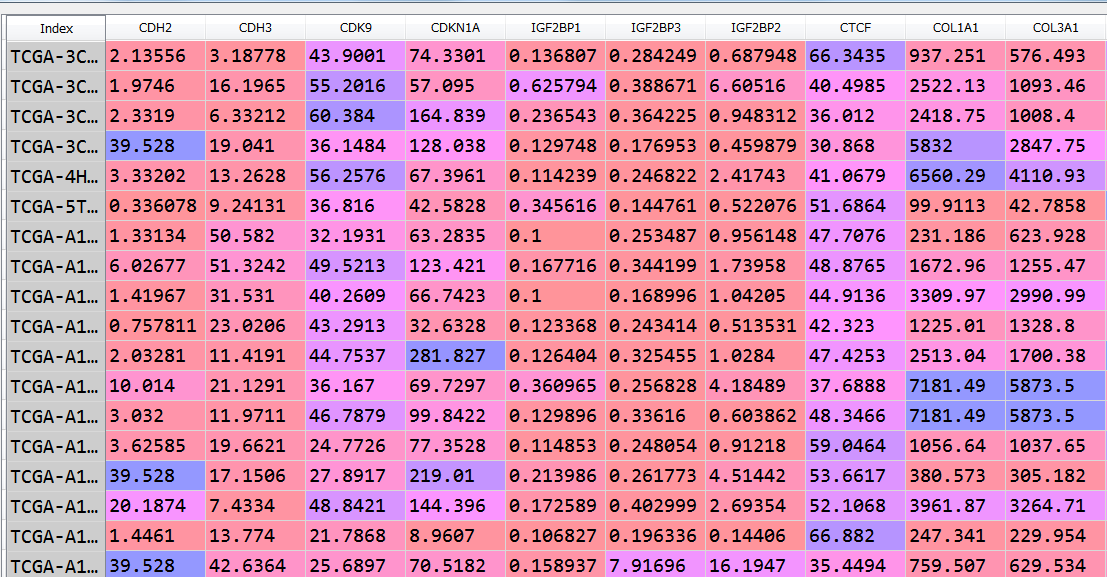
作成したPython解析プログラムファイルの一覧

|  |  |
| --- | --- |
| 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 |

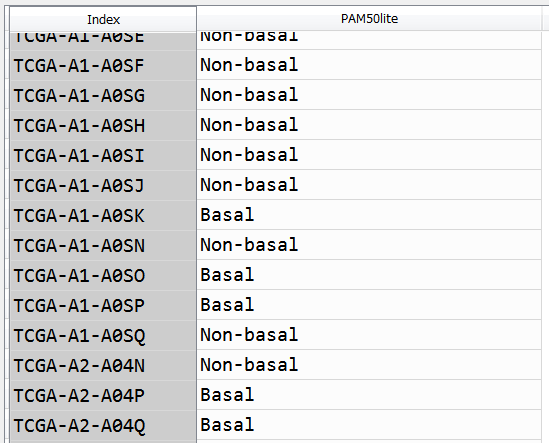
（１）解析対象データの構築

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



患者のクラス値



（２）遺伝子発現データのプロット

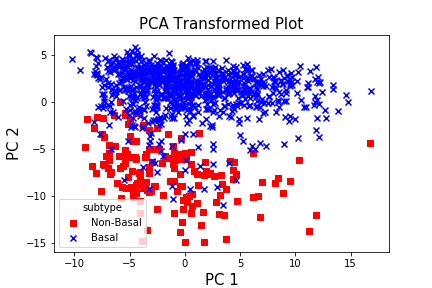
　　　　省略

（３）PCA変換後のデータ

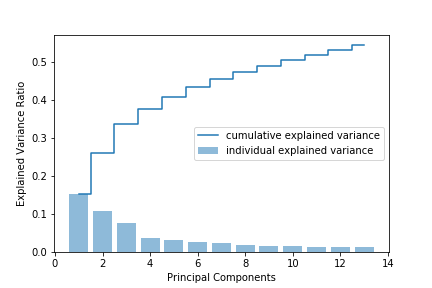
　オリジナル・データに対して主成分分析解析を実施して、第１主成分及び第２主成分をＸ軸、Ｙ軸として散布図を作成したところ、以下のとおりであった。

　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

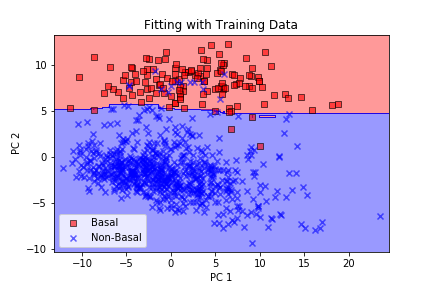
③PCA変換

　モデルパラメータは以下の通りと設定した。

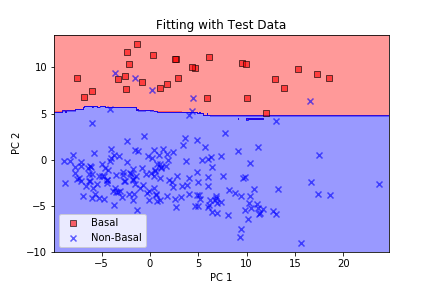
|  |  |  |
| --- | --- | --- |
| パラメーター | 値 |  |
| n\_component | 2 |  |
| n\_estimators | 20 |  |
| max\_depth | 5 |  |

　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。



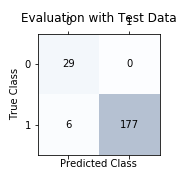
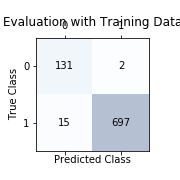
　次に，学習データに基づいて構築されたモデルにより、検証データでの妥当性を確認した。学習データに比べれば正答率は低下しているものの、検証データにおいても、ほぼ適切に分類できていることが示された。



　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。



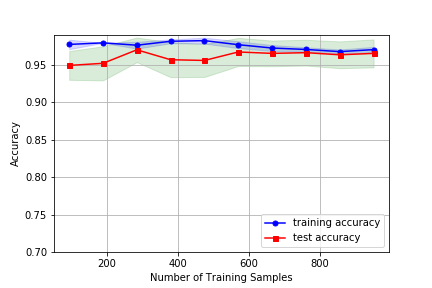
Accuracy with Training Data: 0.980

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。

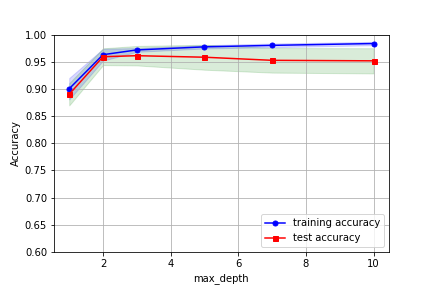


（６）PCA変換後データに対するRandom Forestモデルの検証曲線

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。

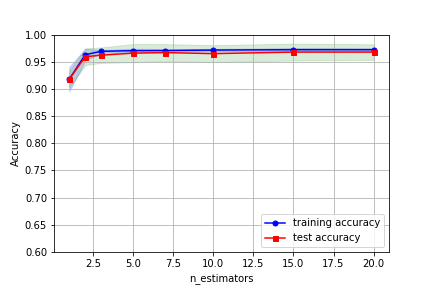
最終モデルとしては、max\_depthは3とする。



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。

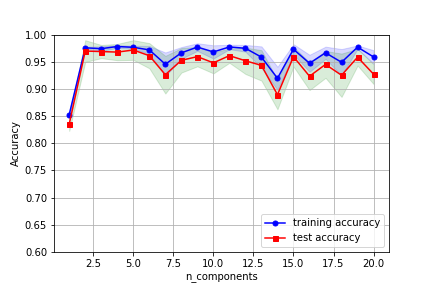


② ＰＣＡのについてのチューニング

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。

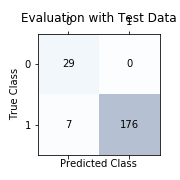
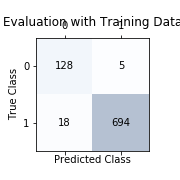
最終モデルパラメータ

|  |  |  |
| --- | --- | --- |
| 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。

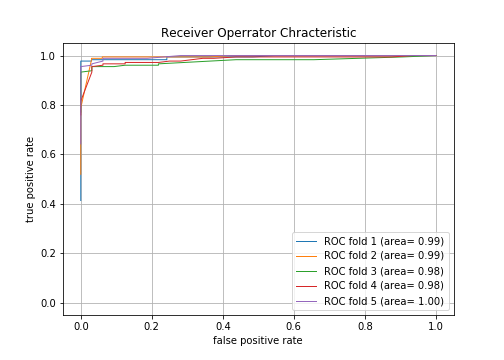


Accuracy with Training Data: 0.973

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