知識発見と特徴選択のための自動化予測モデリング①

～腸内微生物プロファイルを使用した大腸がんの予測診断モデルの作成～

Ioannis Tsamardinos, Paulos Charonyktakis, Georgios Papoutsoglou, Giorgos Borboudakis, Kleanthi Lakiotaki, Jean Claude Zenklusen, Hartmut Juhl, Ekaterini Chatzaki & Vincenzo Lagani

npj Precision Oncology **volume 6**, Article number: 38 (2022)
doi: https://doi.org/10.1038/s41698-022-00274-8

JADBioのAutoMLアプローチは、精密医療やトランスレーショナルリサーチを加速させることができます。具体的には、新たな生物学的知見、精密医療予測モデル、創薬ターゲット、癌やその他の疾患の非侵襲的診断につながる新規バイオシグネチャーやバイオマーカーの発見を促進する可能性があります。

知識発見のための特徴選択

教師あり学習では、特徴選択（変数選択または属性選択）のクラスに属する方法を予測モデリングアルゴリズムと組み合わせることで、（バイオ）シグネチャーを特定することができます。**このような予測シグネチャーを特定することは、知識の発見、分子病態生理学的メカニズムへの洞察の獲得、新規創薬ターゲットの同定、または最小限の測定要件での診断/予後アッセイの設計にとって非常に重要です。**

多くの場合、知識発見のための特徴選択は分析の主要な目標であり、予測モデルは副次的なものに過ぎません。 特徴選択は、差次的発現分析(differential expression analysis)とは異なります。**特徴選択では、特徴の相関を個別ではなく、組み合わせで（多変量で）調べます。**無関係な機能を削除するだけでなく、選択された特徴を考慮して予測に冗長な特徴も削除します。

既存の AutoML ツールの問題点はAutoML ツールは、医療を改善するための独自の機会を提供する方法で、分析と予測モデリングのプロセスをエンドツーエンドで自動化しようとします。これらのツールは、可能な限り最良のモデルを得るために、アルゴリズムとそのハイパーパラメータ値の何千もの組み合わせを自動的に試行します。 ただし、最先端の AutoML ツールは、トランスレーショナル リサーチャーのすべての分析ニーズをカバーできるわけではありません。**まず、特徴選択に焦点を当てていません。原則として、トレーニング データ内のすべての分子量を使用して測定し、予測を提供する必要があるモデルを返すため、解釈、知識発見、および費用対効果の高いベンチトップアッセイに変換する機能が妨げられます。**

第 2 に、モデルのサンプル外の予測性能の信頼できる推定値を提供することに重点を置いていません。後者は、モデルの臨床的有用性を評価する開業医にとって特に重要です。第三に、モデルの解釈、説明、意思決定、臨床での適用に必要なすべての情報を提供していません。

信頼性の高いパフォーマンス推定に焦点を当てると、少なくとも3つの理由から、オミックス解析では特に困難であることに留意する必要があります。

1. 典型的なオミックスデータセットの低サンプル数。生物学的データセットに含まれるサンプルが100 未満であることは珍しくありません。2021 年 10 月の時点で、Gene Expression Omnibus によって提供される 4348 の精選されたデータセットの 74.6% は 20 以下のサンプルとなっています。これは、希少がんや疾患、実験的治療、および測定コストが高い場合によくみられます。サンプルサイズが小さい場合、モデルの統計的検証のために分子プロファイルの大部分を保持する余裕はありません。

2. 複数の機械学習アルゴリズムまたはパイプラインを試してさまざまなモデルを生成し、最もパフォーマンスの高いモデルを選択すると、その予測パフォーマンスが体系的に過大評価されます (バイアス) 。この現象は、統計学では「勝者の呪い」と呼ばれ、機械学習では帰納アルゴリズムにおける多重比較問題と呼ばれています。

オミクスデータセットは、最大数十万の特徴 (次元) を測定します。このような高次元データは、ゲノム、トランスクリプトーム、メタボロミクス、プロテオミクス、コピー数多型、一塩基多型 (SNP) GWAS プロファイリング、および複数のモダリティを含むマルチオミクス データセットなど最新のバイオテクノロジーによって生み出されています。統計学やバイオインフォマティクスで繰り返し指摘されているように、多数の特徴 (p) と低サンプル サイズ (n) の組み合わせ、または「p が大きく、n が小さい」設定は、モデルの過剰適合やパフォーマンスの過大評価の問題を引き起こす有名な難題です。このような課題は、精密医療研究コミュニティでも最近注目されています。

Scope

JADBio は、**auto-sklearn**、**TPOT**、**GAMA**、**AutoPrognosis**、および**Random Forests**と定性的に比較され、トランスレーショナル リサーチャーが意思決定を支援し、モデルの臨床応用に必要な豊富な独自の機能を提供することが実証されています。JADBio の機能を説明および実証するために、大腸がんのマイクロバイオームに関するケーススタディを使用します。JADBio はまた、メタボロミクス、トランスクリプトミクス (マイクロアレイおよび RNA-seq)、およびメチロミクスを測定した、乾癬から癌まで、122 の疾患と対応するコントロールにまたがる 360 の公開生物学的データセットで比較的かつ定量的に評価されます。典型的なオミックスデータセットにおいて、JADBioはわずかな分子量でシグネチャーを同定し、競争力のある予測パフォーマンスを維持することが示されました。同時に、検証のためにサンプルを失うことなく、トレーニング データのみからモデルのパフォーマンスを確実に推定します。対照的に、いくつかの一般的な AutoML パッケージは、モデルのパフォーマンスを大幅に過大評価することが示されています。

Data

575 のサンプルを含む 5 つのコホート (大腸がん285例、健常対照290例)*Wirbel, J. et al.* **Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer.** Nat. Med. 25, 679–689 (2019). https://www.nature.com/articles/s41591-019-0406-6

Tasks/Analyses

分析タスクは次のとおりです。
a.腸内微生物プロファイルから、大腸がんの予測・診断モデルを作成
b.モデルの予測性能を推定する。
c.大腸がんを予測する微生物種のバイオシグネチャーを同定する、および
d.大腸がんモデルを新しいデータに適用する。特に、各コホートを予測モデルの構築に使用し、他のすべてのコホートをその外部検証セットとして使用します。

結果

a. 大腸がんマイクロバイオームデータの解析

詳細な結果セットをTable 1に示します。全体として、JADBio分析は原著論文と同様のパフォーマンスを提供し、原著論文とJADBio分析の最大差は0.08 AUCポイントでした。ただし、JADBioは最小限の人的労力で、平均してより少ない予測機能 (最大 25) を選択することに注意してください。*Wirbel*らはモデルのアンサンブルアプローチを採用しているため、各モデルに使用される特徴量の数を抽出することはできませんが、実用上、彼らのモデルは849の特徴量一式を使用しています。測定されたすべてのバイオマーカーをモデルに使用すると、その解析の知識発見の側面や診断アッセイの設計への応用が制限されます。

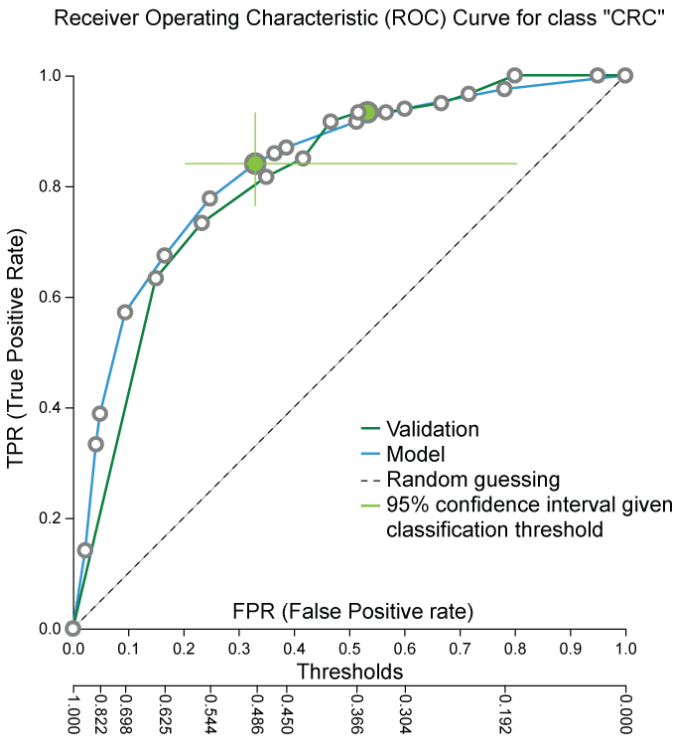
腸内細菌プロファイルを用いた大腸がんの診断モデルインスタンス

Fig. 1aは、CNコホートで最も性能の良いモデルを生成する構成を報告したものです。非線形モデル、具体的には1000種類の決定木を含むRandom Forest が生成されます。この分析の基準として「解釈可能性（Interpretability）」が選択された場合、最も解釈可能なモデルは線形リッジ回帰型のモデル (Fig. 1b) であり、その標準化係数はFig. 1c に示されています。

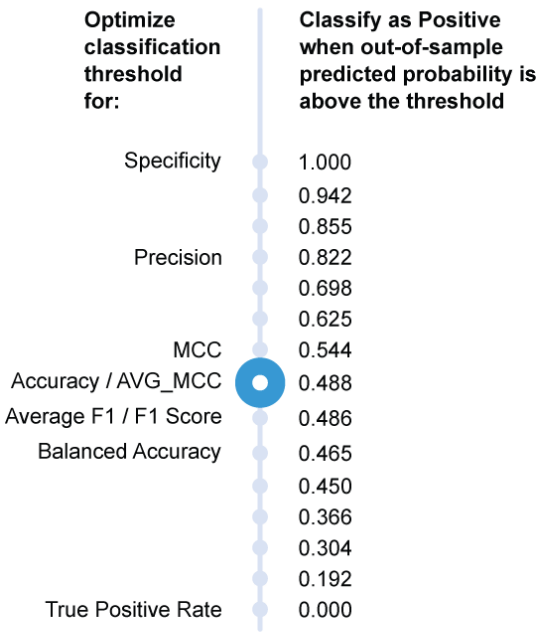
Fig. 1 中国 (CN) コホートでトレーニングされた、最高のパフォーマンスと最高の解釈可能なモデル。
a. すべてのトレーニング データに適用した時に、最終的に最高のパフォーマンスを発揮するモデルの生成につながる Winning 構成。これは定数値特徴の削除と全特徴の標準化を行った後、全特徴をモデルに含め（FullSelector、すなわち特徴選択なし）、Random Forestアルゴリズムでモデル化することにより生成されます。使用するハイパーパラメーター値も表示されます。b. 最良の解釈可能なモデルの生成につながる構成。これは線形線形リッジロジスティック回帰型のモデルで、特徴選択のためのLasso回帰が先行します。c. 解釈可能なモデルは、最高のパフォーマンスを発揮するモデルとは対照的に視覚化できます。 JADBio は、選択された各特徴と切片項のモデルの標準化された係数を示します。

b.大腸がんモデルの予測性能の評価

a



b



c

	Validation	Train			
Metric	Estimate	Mean estimate	95% confidence interval	Unadjusted estimate	Base line
Area Under The Curve	0.834	0.845	[0.771, 0.906]	0.862	0.500

Fig. 2 中国（CN）コホートで学習し、ドイツ（DE）コホートで検証した、Feature selectionに最適化されたwinning modelの予測性能。
a.トレーニング セットで推定され、複数の異なる構成を試行するために制御されたROC（Receiver Operating Characteristic） 曲線は、青い線で示されています。これは、すべての異なる分類しきい値 (下のx軸) について、偽陽性率 (上のx軸) と真陽性率 (y軸) の間のすべてのトレードオフを示しています。円をクリックすると、対応するしきい値が選択されます。ユーザーは、パフォーマンスの測定基準がどのように影響を受けるかを確認でき、臨床的に最適なしきい値の選択が容易になります。緑色の十字は、ROC上のその点の各次元における信頼区間を示しています。外部検証セットとして使用したDEコホートに対するモデルの予測値が達成したROC曲線は、緑色の線で示されています。トレーニングから推定された青い線に忠実に沿っています。b. JADBioが提案するさまざまなしきい値と、それに対応する最適化された性能の指標の一覧です。例えば、accuracyは閾値0.488に最適化され、balanced accuracyは0.465に最適化されます。c.トレーニングによって推定され (太字の青)、バリデーションで達成された (太字の緑) の、しきい値に依存しないROC Area Under the Curve (AUC) が、その信頼区間とともに報告されます。その未調整の推定 (つまり、複数の構成を試すために調整されていない) も報告されており、平均して楽観的であると予想されます。 JADBioが計算したaccuracy, precision, recall (図には示されていない) など他のすべての推定値も、同様に複数の構成を試すために調整されています。

c.大腸がんを予測する微生物種のバイオシグネチャーの同定

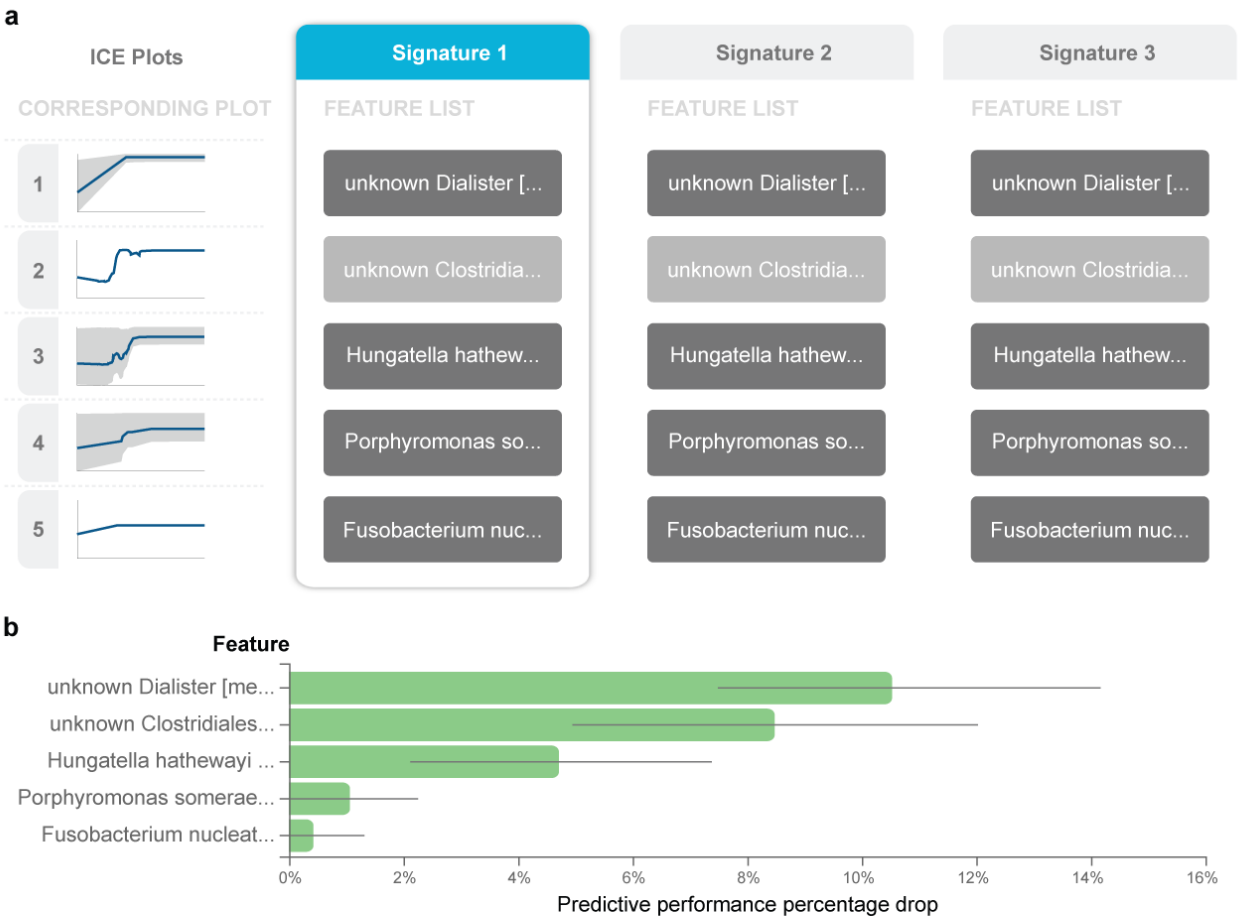


Fig. 3　フランス (FR) コホートをトレーニング データとして使用した、最も優れたモデルの特徴選択 (バイオシグネチャー検出) の結果①
a. 3つのバイオシグネチャーが同定され、それぞれが5つの特徴を含んでいます。各シグネチャーは最適な予測モデルを導き出します。 4 つの特徴は共通していますが、2 番目の特徴はシグネチャーによって異なります。b. 各特徴がシグネチャーから削除され、残りが保持された場合のパフォーマンスの相対的な低下を評価する、特徴の重要度プロット。

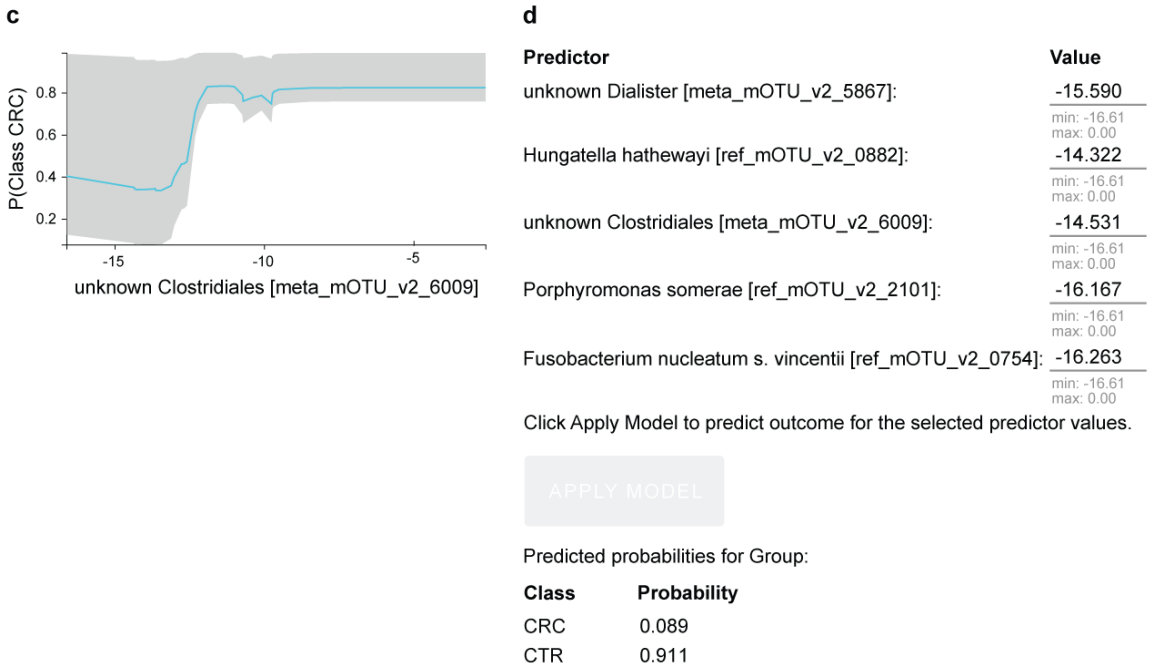


Fig. 3　フランス (FR) コホートをトレーニング データとして使用した、最も優れたモデルの特徴選択 (バイオシグネチャー検出) の結果②
c. unknown Clostridiales [meta_mOTU_v2_6009]に対する Individual Conditional Expectation (ICE) プロット: 存在量が多いほど、大腸がんの可能性が高くなります。 傾向は非線形で非単調 (ステップ関数に近い) であることに注意してください。d. 架空の新しいサンプルの予測を取得するためのモデルの手動適用が示されています。値は対数変換された相対濃度であり、元の数値が対数の底よりも小さい場合は負になる可能性があります。

JADBio は、知識発見を容易にするために、モデリングと同時に特徴選択 (バイオシグネチャーの発見) を実行します。JADBio は、複数の特徴選択を実行します。つまり、存在する場合、統計的に同等になるまで、同等に予測可能なモデルにつながる複数の代替特徴サブセットを識別することができます。このコホートで最も優れたモデルは、3つのバイオシグネチャー（Fig. 3a）を示しており、それぞれ、合計849の特徴の中からわずか5つの特徴（すなわち、微生物種）を選択しています。すべてのシグネチャーの最初の特徴は、「unknown Dialister」と名付けられ、Dialister属の未知の微生物の相対的な存在量です。2 番目の特徴は、クロストリジウム目に属する未同定の 3 種類の微生物「unknown Clostridiales [meta_mOTU_v2_6009]」、「unknown Clostridiales [meta_mOTU_v2_5514]」、「unknown Clostridiales [meta_mOTU_v2_7337]」間でシグネチャーが異なることです（フルネームはFig. 3a で省略されています）。つまり、これら3つの微生物は、それぞれ他の微生物に置き換わることができ、同じように予測できるモデルであるため、この結果の予測に関しては情報的に等価です。

興味深いことに、未知の Dialister と Hungatella hathewayi の微生物は、結果 (両側 t 検定) に最も関連する (ペアワイズ) マーカーの中でそれぞれ 1 位と 2 位にランク付けされますが、Porphyromonas somerae と Fusobacterium nucleatum (単変量、無条件) の p 値は それぞれ27位と46位でした。この例は、遺伝子を個別に調べる標準的な発現差異解析と、バイオマーカーを一緒に調べる特徴選択との違いを端的に示しています。多変量解析では、結果の予測に向けて互いに補完し合うため、結果とのペアワイズ関連が比較的弱い特徴が選択される場合があります。

d.大腸がんモデルを新たな未公開データに適用する

ユーザーは、外部検証のために新しいラベル付きデータセットにバッチ形式でモデルを適用したり、予測値を得るためにラベルのないデータセットに適用したり、あるいはモデルを実行ファイルとしてダウンロードして自分のコードに埋め込むことができます。

JADBio は、異なる集団間で伝達される腸内微生物データから、大腸がんの正確な診断モデルを作成することに成功しました。 JADBio は、利用可能な 849 個の特徴のうち、最大 25 個の特徴のシグネチャーを識別しました。 モデルの構築にはコーディングは必要ありませんでした。

上記の事例では、FRコホート（Fig.3a）において、それぞれ5種類の微生物種からなる3つの特徴サブセットが同定され、同等の予測モデルにつながりました。それぞれ、予測目的には十分です。しかし、特徴選択を知識発見のために用いる場合、3 つのシグネチャーをすべて報告しないのは誤解を招きます。また、診断用アッセイの設計者にとっては、これら3つのシグネチャーをすべて特定することは、設計上の選択肢となります。複数の特徴選択問題は、これまでコミュニティではほとんど注目されておらず、利用可能なアルゴリズムもほんの一握りでした。JADBioは、この種の機能を提供する唯一のツールです。

【お問い合わせ先】
フィルジェン株式会社　バイオインフォマティクス部
TEL：　052-624-4388 (9:00～17：00)
FAX：　052-624-4389
E-mail：　biosupport@filgen.jp