その人工知能は本当に信頼できるのか？

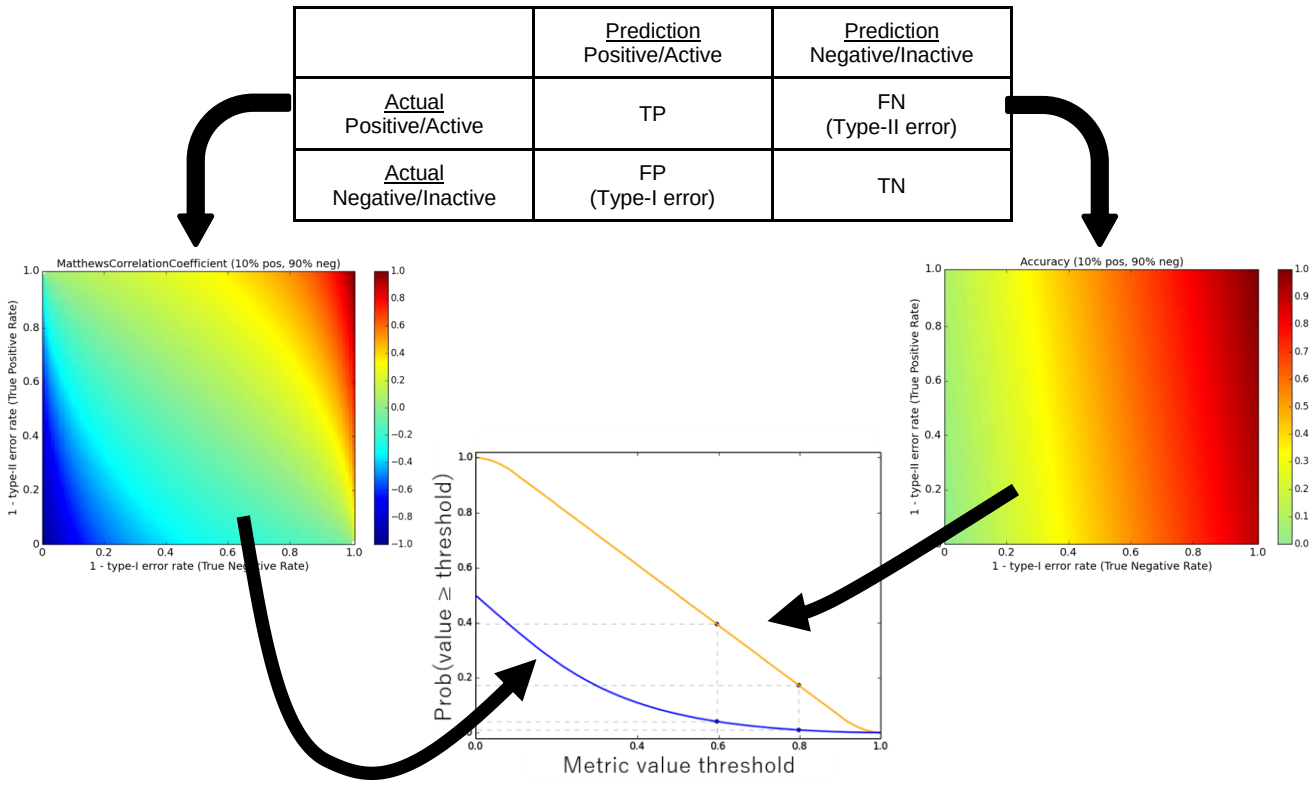
―人工知能の性能を正確に評価する方法を開発―

概要

人工知能（AI）によるビッグデータ解析は、医療現場や市場分析など社会のさまざまな分野での活用が進み、今後さらなる普及が予想されています。また、創薬研究などで分子モデルの有効性を予測する場合にも、AI は主要な検証手段として重視されています。ところが、私たちは肝心の AI の性能を正しく評価できているのでしょうか？

J.B.Brown 京都大学大学院医学研究科 講師は、ヒートマップ（可視化グラフ）を用いた統計学的分析によって、AI の性能評価指標そのものの有効性を網羅的に検証し、分野を問わず正確に AI の性能を評価できる手法を世界で初めて開発しました。信頼性の高い AI の開発に加えて、ビッグデータを用いた創薬研究や治療法の創出などに貢献することが期待されます。

本研究は、米国の科学誌「*Molecular Informatics*」に 2018 年 2 月 14 日付で掲載されました。



AI の性能評価指標は、ヒートマップと iCDF（逆累積分布関数）を使って検証することができる。ACC が AI の性能を過大に評価する危険性がある一方、MCC は性能を正確に評価できる厳密な評価指標であることなどが分かる。実験で使う AI を評価する前に、本研究の手法によって指標そのものの特性を十分に吟味すべきである。

１．背景

AI によるビッグデータ解析は、市場分析や金融機関におけるローン滞納調査など社会のさまざまな分野での活用が広がっています。同様に創薬研究・医療現場のスクリーニング検査においても、AI を使用したコンピューターモデルの二項分類による解析が主要な研究／検査手法となっています。このように AI が社会に普及するにあたっては、その性能を正しく評価することがきわめて重要です。

二項分類モデルでは、「はい（陽性・真）」と「いいえ（陰性・偽）」でデータを分類し、TP（True Positive：正しく陽性と分類）・FN（False Negative：誤って陰性と分類）・FP（False Positive：誤って陽性と分類）・TN（True Negative：正しく陰性と分類）という 4 種類の結果が得られます。AI の性能は、データをこの 4 種類に正しく分類できた割合によって、さまざまな統計的指標を用いて評価されてきました。

しかし、例えば特定の分子を検出する場合に、実験における検出成功率が、コンピューターモデルによる事前予測を大きく下回るという事例がしばしば報告されています。その根本的な原因は、コンピューターモデルすなわち AI の性能を過大に評価した統計的指標にあると考えられます。これまでは、AI の性能評価指標として TPR（True Positive Rate：真陽性率）と ACC（Accuracy：正確率）をはじめとする数種類の指標が用いられてきましたが、これらの指標は本当に AI の性能を正しく評価できていたのでしょうか？

２．研究手法・成果

本研究は、上記の課題を解決するために、AI の性能を統計的指標によって正確に評価する手法を開発しました。この手法は以下のように、TPR や ACC など各指標の特性と有効性を、ヒートマップ（可視化グラフ）と iCDF（Inverse Cumulative Distribution Function：逆累積分布関数）を使った統計学的な解析によって検証するものです。

本研究では、二項分類モデルを評価する指標として、上記の TPR と ACC に加えて、BA（Balanced Accuracy：平均正解率）・PPV（Positive Predictive Rate：陽性的中率）・F1 値（F1 Score：PPV と TPR の調和平均）・TNR（True Negative Rate：真陰性率）および MCC（Matthews Correlation Coefficient：マシューズ相関係数）を検証の対象としました。各指標が取りうる値は、MCC は-1 から+1 まで、その他は 0 から+1 までとなります。

まず、AI に陽性と陰性のバランスが取れたデータ（陽性 50%・陰性 50%）と、陽性と陰性のバランスが極端に崩れたデータ（陽性 10%・陰性 90%）とを与えた場合に、ACC と MCC が下した性能評価についてヒートマップを作成して比較しました（図１：赤みが強まるほど評価が高いことを示す）。その結果、MCC が AI の性能を厳密に評価するのに対して、ACC は過大に評価する可能性が高く、この性質はバランスの崩れたデータセットではより顕著に現れることが分かりました。具体的には、MCC が陽性と陰性のどちらも正しく判定した場合でなければ 0.6 以上の高い評価を下さない一方で、ACC は陽性をひとつも正しく分類できない AI に対しても高評価を与えてしまうのです。

次に、ACC と MCC について、iCDF を使って特定の評価を得られる確率を求めました（図２）。MCC ではバランスの取れたデータセットでも 0.6 以上の高評価を得られる確率は 10%以下と低く、極端にバランスの崩れたデータセットではさらに確率は低下します。一方 ACC では、0.6 以上の高評価を得られる確率が高いことに加えて、バランスの崩れたデータセットではむしろその確率が上昇してしまうことが分かりました。こ

のように、ACC は AI の性能を過剰に評価する危険性が高く、AI を評価する場合には、より厳密な指標である MCC を使う方が望ましいといえます。

続けて、その他の４つの指標についても、同様にヒートマップ（図 3・図 4）と iCDF（図 5）によって特性を評価しました。バランスの取れたデータでは、F1 値は陽性を正しく分類する AI の性能を過剰に評価する危険性が高く、BA は ACC と同様の傾向を示しました。一方バランスの崩れたデータでは、TNR は ACC との相関性が確認されるため、ACC と同様に使用には注意が必要であることが分かりました。F1 値と PPV については、データのバランスが崩れた影響で高評価を与える範囲が縮小するため、この場合には MCC のように厳密な指標として AI の性能評価に有効であることを示しています。iCDF によっても、ヒートマップで示された各指標の同様な性質を確認することができました。なお、ヒートマップと iCDF はどちらも、どのようなバランスのデータに対しても適用できるため、検証に用いるデータセットのバランスに応じて評価指標の特性を把握することができます。

本研究ではさらに、ROC（Receiver Operating Characteristic：受信者動作特性）曲線と、ROC 曲線下部の面積 AUC（Area Under Curve）を用いた AI の性能評価方法についても検証しました。その結果、この AUC を用いた評価方法は MCC や F1 値といった評価指標との相関性が無く、事前に陽性と陰性が判明しているデータセットに合わせて設計された AI の評価には使えるものの、実証実験で陽性と陰性を分類する AI の性能評価には使えない、という欠陥があることが明らかになりました。

以上のことから、AI の性能評価指標の中には ACC のように性能を過大に評価するものがあるため、AI を使ってデータ分類を行う場合には、本研究で行ったように、実験を行う前にヒートマップと iCDF によって評価指標そのものの特性を十分に吟味するべきであることが分かりました。今回の実験によって示したとおり、社会に浸透しつつある AI も、その情報の正確性を評価した上で有効利用することが必要不可欠といえます。

３．波及効果、今後の予定

本研究は、AI の性能評価指標そのものの有効性を、ヒートマップと iCDF を使って統計学的に検証した世界で初めての成果で、創薬スクリーニングやケミカルバイオロジーに限らず、どの分野の AI に対しても適用できる画期的なものです。また、成果をただちに活用できるように、論文の追加データとしてヒートマップと iCDF を作成するプログラムを公開しました。どのようなデータセットに対しても、実証実験で正確な分類ができる「堅牢な」AI の開発に貢献することが期待されます。

＜論文タイトルと著者＞

タイトル：Classifiers and their Metrics Quantified

著者：J.B.Brown

掲載誌：*Molecular Informatics*　DOI：10.1002/minf.201700127（オープンアクセス）

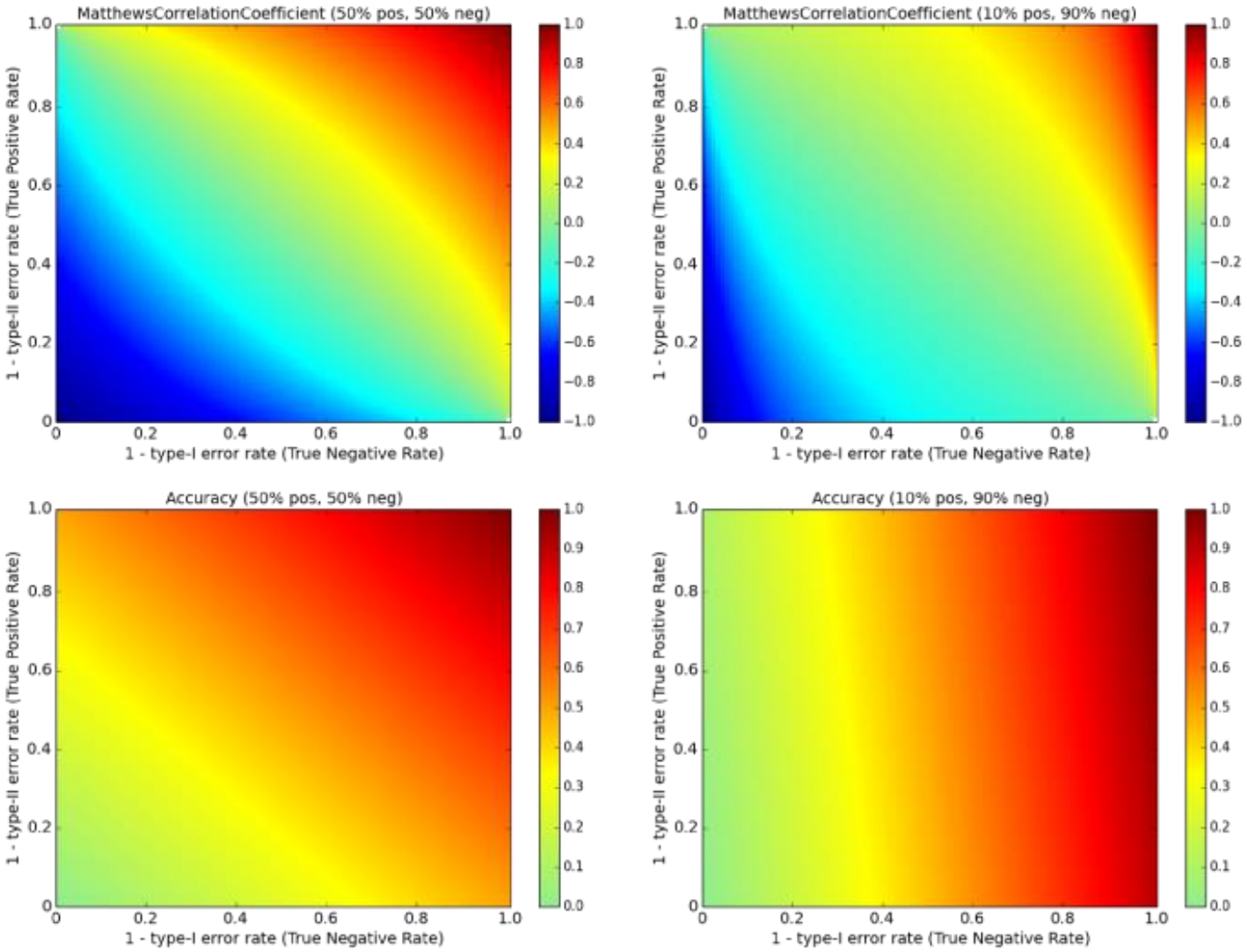


図１：MCC と ACC の AI 性能評価ヒートマップ。左側がバランスの取れたデータセット、右側がバランスの崩れたデータセットについて図示したもの。縦軸に TPR（True Positive Rate：陽性を正しく分類した確率）、横軸に TNR（True Negative Rate：陰性を正しく分類した確率）を取っている。ACC は 0～1、MCC は-1～1 の範囲で結果が出る。値が高いほど高評価となる。MCC が ACC に比べて高評価の出にくい厳しい評価指標であることが視覚的に確認できる。

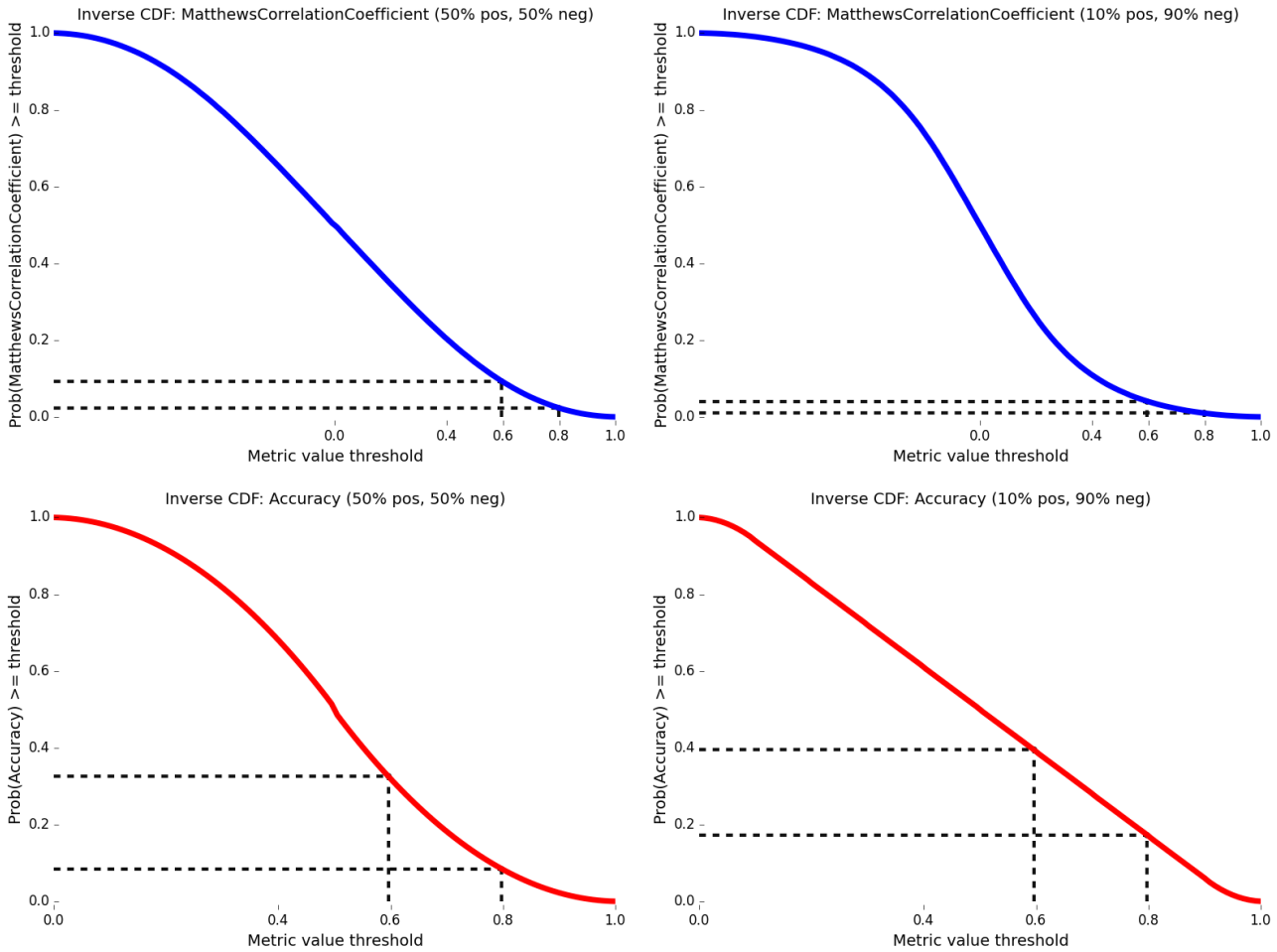


図２. iCDF を使って、特定の評価を得られる確率を ACC と MCC について求めてグラフにしたもの。MCC では 0.6 以上の高評価を得られる確率が低く、極端にバランスの崩れたデータセットではさらに確率が低くなる。一方 ACC では、もともと高評価を得られる確率が高いことに加えて、バランスの崩れたデータセットではむしろその確率が上昇してしまうことが分かる。

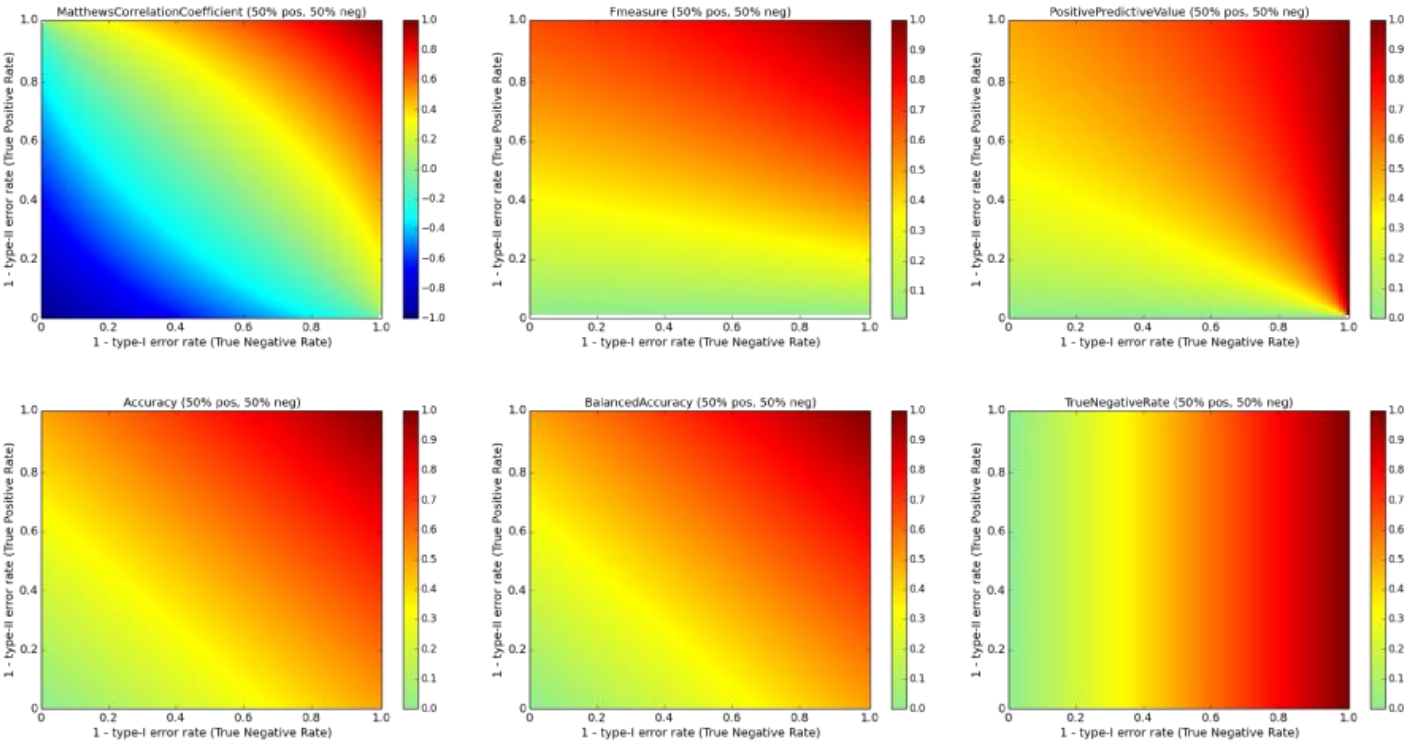


図３：６つの評価指標について、バランスの取れたデータ（陽性 50%・陰性 50%）で作成したヒートマップ。上段は左から MCC、F1 値、PPV、下段も同じく左から ACC、BA、TNR。

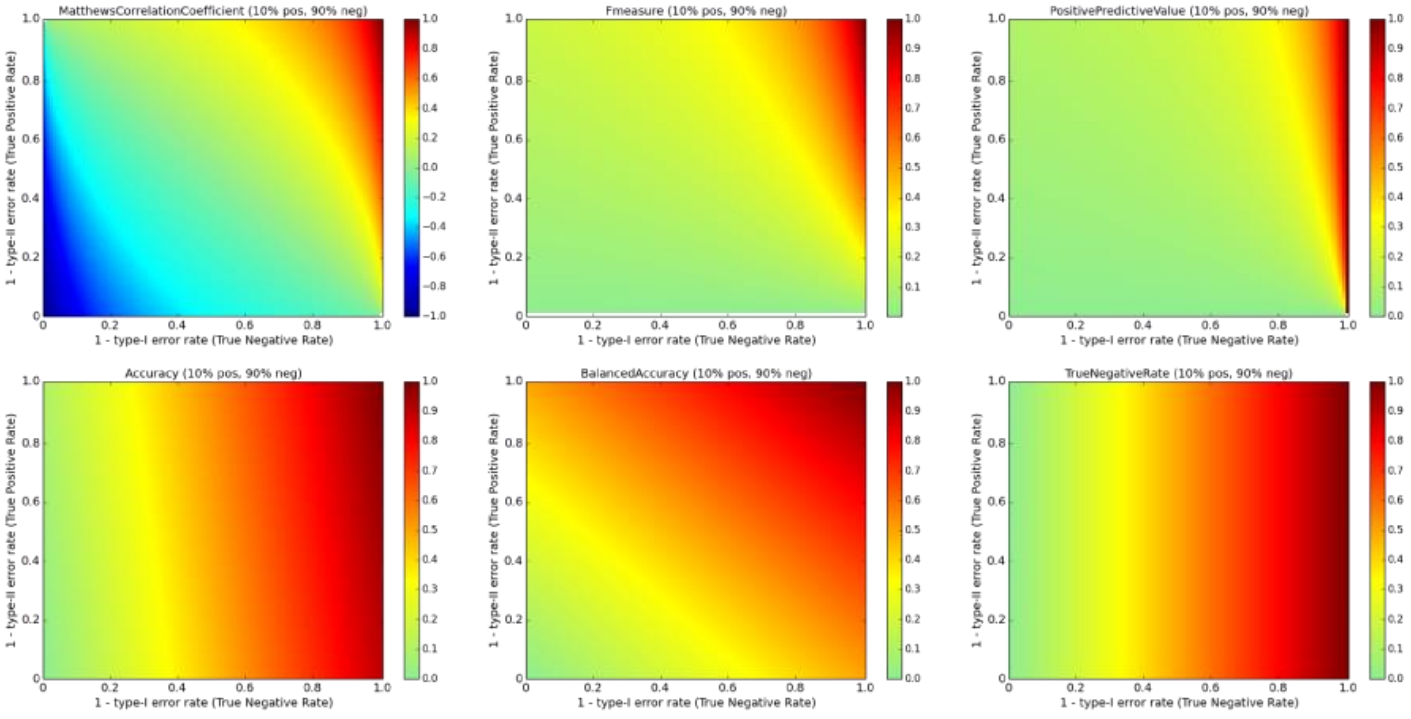


図４：6 つの評価指標について、バランスの崩れたデータ（陽性 10%・陰性 90%）で作成したヒートマップ。並び方は図３と同じ。

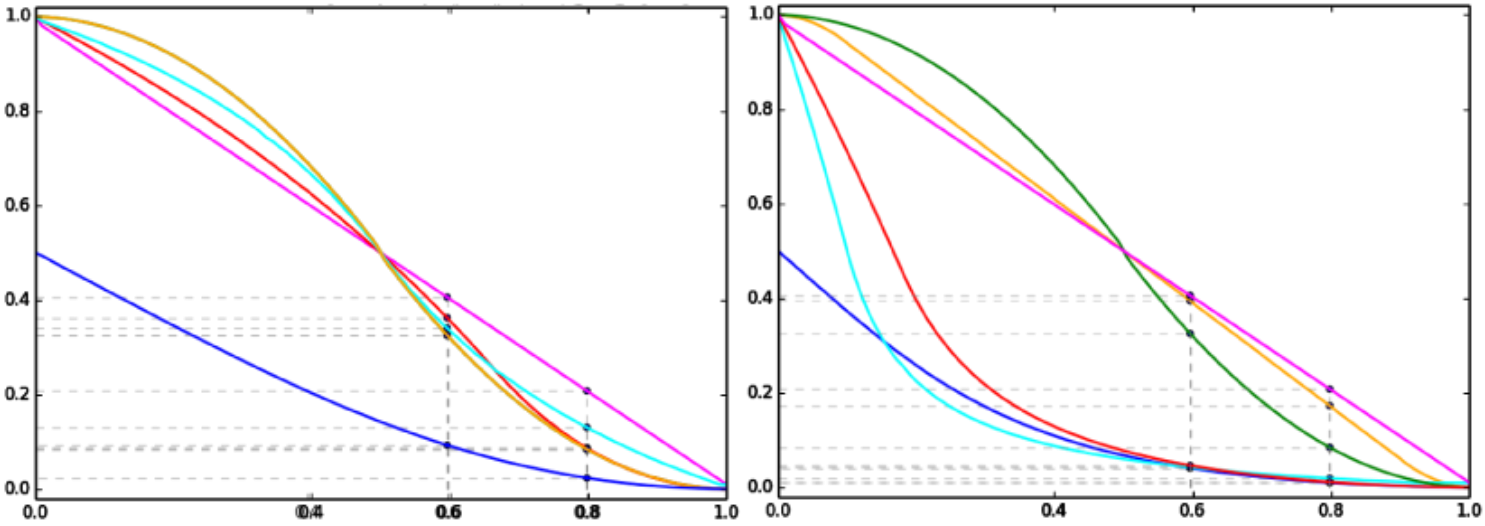


図 5 : 6 つの評価指標について、iCDF によって解析したグラフ。左がバランスの取れたデータ（陽性 50%・陰性 50%）の場合、右がバランスの崩れたデータ（陽性 10%・陰性 90%）の場合。グラフの色はそれぞれ青＝MCC、水色＝PPV、赤＝F1、オレンジ＝ACC、緑＝BA、紫＝TNR。左図では ACC と BA が重なっている。