Classification assessment methods

備考

著者

Alaa Tharwat

掲載

''Classification assessment methods," {\itApplied Computing and Informatics}, Vol. 16, No. 2, 25 pages, 2020.

Abstract

分類技術は、様々な科学分野で多くの応用がなされています。分類アルゴリズムの評価にはいくつかの方法があります。このような指標の分析とその重要性は、異なる学習アルゴリズムを評価するために正しく解釈されなければならない。これらの尺度の多くはスカラーメトリクスであり、その中にはグラフィカルな手法を用いたものもある。本論文では、この分野に関心のある研究者のための総合的な資料となることを目的として、分類評価指標の詳細な概要を紹介します。この概説は、まず、2値分類問題や多クラス分類問題における混同行列の定義を強調することから始まります。また、多くの分類尺度についても詳細に説明し、各尺度におけるバランスのとれたデータと不均衡なデータの影響を示す。例示的な例が紹介され、(1)2値分類問題や多値分類問題におけるこれらの尺度の計算方法、(2)バランスのとれたデータや不均衡なデータに対するいくつかの尺度のロバスト性を示しています。さらに、受信機動作特性(ROC)、精度-リコール、検出誤差トレードオフ(DET)曲線のようないくつかのグラフィカルな測定値が詳細に示されています。さらに、ROC,PR,DET曲線をプロットする前処理の手順を説明するために、ステップバイステップのアプローチで、異なる数値例を示している。

2. Classification

評価方法は、分類器の性能を評価し、分類器のモデル化を導く上で重要な要素です。分類プロセスには大きく分けて、訓練段階、検証段階、テスト段階の3つの段階があります。モデルは入力パターンを用いて訓練され、このフェーズを訓練フェーズと呼ぶ。これらの入力パターンは、モデルを訓練するために使用される訓練データと呼ばれます。この段階では、分類モデルのパラメータが調整されます。訓練誤差は、訓練されたモデルが訓練データにどれだけ適合するかを測定します。しかし、訓練されたモデルが訓練段階で使用される同じデータに適合するため、訓練誤差は常にテスト誤差や検証誤差よりも小さくなります。学習アルゴリズムの目標は、学習データから学習して、目に見えないデータのクラスラベルを予測することです。しかし、テストサンプルのクラスラベルや出力が未知であるため、テスト誤差やサンプル外誤差を推定することはできません。これが，検証フェーズが訓練されたモデルの性能を評価するために使用される理由です．検証段階では，検証データは，モデルのハイパーパラメタを調整しながら，訓練されたモデルの偏りのない評価を提供する．

クラスの数に応じて、クラスが2つしかない**2クラス分類**と、**クラスの数が2つより多い多クラス分類**の2種類の分類問題があります。2つのクラス、すなわち、正のクラスをP、負のクラスをNとする2値分類があるとします。未知のサンプルは、PまたはNに分類され、学習段階で学習された分類モデルは、未知のサンプル

の真のクラスを予測するために使用されます。この分類モデルは、連続または離散の出力を生成します。分類モデルから生成される離散出力は、未知/テストサンプルの予測された離散クラス・ラベルを表し、連続出力はサンプルのクラス・メンバーシップ確率の推定を表します。

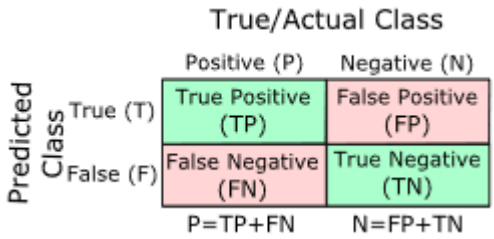


図1:2x2混同行列の例を示します。2つの真のクラスPとNがあり，予測されたクラスの出力は真か偽かである.

[図1]は、2x2混同行列または分割表の要素を表す4つの可能な出力があることを示しています。 **緑の対角線は正しい予測値を、ピンクの対角線は正しくない予測値を表しています。** サンプルが陽性であり，陽性に分類された場合，すなわち正しく分類された陽性サンプルは，真の陽性(TP)としてカウントされ，陰性に分類された場合は，偽陰性(FN)またはタイプIIエラーとみなされます．サンプルが陰性であり、陰性に分類された場合は真陰性(TN)とみなされ、陽性に分類された場合は偽陽性(FP)、誤報、またはタイプIエラーとみなされます。次のセクションで紹介するように、混同行列は、多くの一般的な分類メトリックを計算するために使用されます。

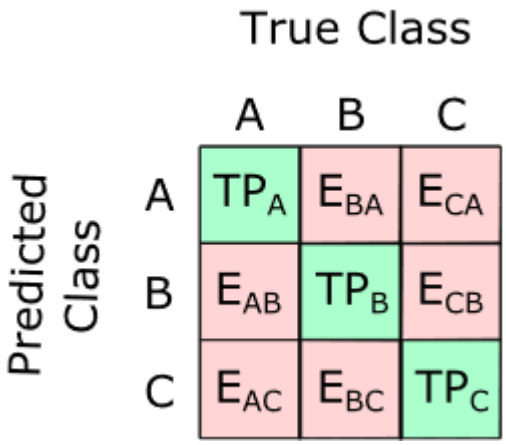


図2：マルチクラス分類テストのための混同行列の例。

図2は、3つのクラス（A,B,C）を持つマルチクラス分類問題の混同行列を示している。このように、TPAはクラスAの真の陽性サンプル数、すなわちクラスAから正しく分類されたサンプル数であり、EABはクラスAからクラスBとして誤って分類されたサンプル数、すなわち誤分類されたサンプル数である。したがって、Aクラスの偽陰性(FN_A)はE_{AB}とE_{AC}の和 $(FN_A = E_{AB} + E_{AC})$ であり、これはクラスBまたはCとして誤って分類されたすべてのクラスAのサンプルの和を示している。一方、行に位置する予測されたクラスの偽陽性FNは、その行のすべてのエラーの合計を表します。例えば、クラスA (FP_A) の偽陽性は次のように計算されます $(FP_A = E_{BA} + E_{CA})$ 。$m \times m$の混同行列では、m個の正しい分類と$m^2 - m$の誤りの可能性のあるがある[22]。

2.1. Classification metrics with imbalanced data.

あるデータセットのあるクラスのサンプル数が他のクラスのサンプル数を上回っている場合には，異なる評価方法が不均衡なデータに敏感に反応します[25]．これを説明するために，図1の混同行列を考えてみましょう．クラス分布は、正のサンプルと負のサンプルの間の比率 $\frac{P}{N}$ で、左列と右列の間の関係を表します。 **両方の列からの値を使用する評価メトリックは、[8]で報告されているように、不均衡なデータに敏感になります。** 例えば，accuracy や precision のような評価指標の中には，混同行列の両列の値を使用しているものがあります．したがって，**このようなメトリクスでは，異なるクラスからの修正ラベルの数を区別することはできません[11]。** 幾何平均（GM）やユーデンの指数（YI）のように、両方の列の値を使用するメトリクスがあり、これらのメトリクスはバランスの取れたデータと不均衡なデータで使用できるので、この事実は部分的に正しいです。これは、片方の列の値を用いるメトリクスは、クラス分布の変化を打ち消すと解釈することができる。しかし、両カラムの値を用いるメトリクスの中には、クラス分布の変化が相殺されてしまうため、不均衡データの影響を受けないものもある。例えば、精度は$Acc = \frac{TP+TN}{TP+TN+FP+FN}$、$GM$は以下のように定義される。

$$GM = \sqrt{TPR \times TNR} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}}$$

したがって，両方のメトリクスは混同行列の両列からの値を使用します．クラス分布の変更は，ネガティブクラス/ポジティブクラスのサンプル数を増減させることで得ることができる．同じ分類性能で、負のクラスのサンプル数をα 倍に増やしたと仮定すると、TNとFPの値はそれぞれαTNとαFPになり、精度は次のようになります。

$$ACC = \frac{TP + \alpha TN}{TP + \alpha TN + \alpha FP + FN} = \frac{TP + TN}{TP + TN + FP + FN}$$

つまり、クラス分布の変化によって精度が左右されることになります。一方、ＧＭ メトリックは

GM = \sqrt{\frac{TP}{TP+FN} \times \frac{\alpha TN}{\alpha TN +\alpha FP}}

\sqrt{\frac{TP}{TP+FN}\times \frac{TN}{TN+FP}}

となり、負のクラスの変化は互いに打ち消し合うことになります。これがGMメトリックが不均衡データに適している理由です。同様に、どのようなメトリックでも、それが不均衡なデータに対して敏感かどうかを調べることができます。

2.2. Accuracy and error rate

精度（Acc）は、分類性能の指標として最も一般的に用いられるものの一つであり、以下のように、正しく分類されたサンプル数と総サンプル数の比として定義されている[20]。

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

ここで、 ＰおよびＮはそれぞれ陽性サンプル数および陰性サンプル数を示す。

精度指標の補数は、エラー率(ERR)または誤分類率である。このメトリックは、正と負の両方のクラスからの誤分類されたサンプルの数を表し、次のように計算されます、$EER = 1 - Acc = \frac{FP+FN}{TP+TN+FP+FN}$[4]。精度とエラー率の両方のメトリクスは、不均衡なデータに対して敏感である。精度に関するもう1つの問題

は，2つの分類器が同じ精度を得ることができても，それらが提供する正誤判定の種類に関して異なる性能を発揮することである[9]．しかし，Takaya SaitoとMarc Rehmsmeierは，彼らの例でバランスのとれたデータと不均衡なデータの精度値が同じであることを発見したため，精度は不均衡なデータに適していると報告しています[17]．彼らの例で精度値が同一であった理由は、バランスデータと不均衡データの TP と TN の和が同一であったためである。

2.3. Sensitivity and specificity

感度、真陽性率(TPR)、ヒット率、またはリコール は、分類器の正のサンプルの総数に対する正に分類されたサンプルの割合を表しており、式（2）[20]に従って推定される。一方、**特異度、真陰性率(TNR)、または逆リコール**は、式(2) [20]のように、陰性サンプルの総数に対する正しく分類された陰性サンプルの比率で表されます。したがって、特異度は正しく分類された陰性検体の割合を表し、感度は正しく分類された陽性検体の割合を表す。一般的に、感度と特異度は2種類の精度と考えることができ、第1は実際の陽性サンプルに対するものであり、第2は実際の陰性サンプルに対するものである。感度は混同行列の同じ列にあるTPとFNに依存し，同様に特異度は同じ列にあるTNとFPに依存する．

$$TPR = \frac{TP}{TP + FN} = \frac{TP}{P}$$

$$TNR = \frac{TN}{FP + TN} = \frac{TN}{N}$$

また、精度は、感度と特異度の観点から以下のように定義することができる[20]。

$$ ACC = \frac{TP+TN}{TP+TN+FP+FN} \ = TPR \times \frac{P}{P+N}+TNR \times \frac{N}{P+N} \ =\frac{TP}{TP+EN}\frac{P}{P+N}+\frac{TN}{TN+FP}\frac{N}{P+N}\ =\frac{TP}{P+N}+\frac{TN}{P+N} \ =\frac{TP+TN}{TP+TN+FP+FN} $$

2.4. False positive and false negative rates

偽陽性率（FPR）は偽警報率（FAR）またはフォールアウトとも呼ばれ、陰性サンプルの総数に対する不正確に分類された陰性サンプルの比率を表します[16]。言い換えれば、不正確に分類された陰性サンプルの割合です。したがって、式(4)の特異度を補完するものである[21]。偽陰性率（FNR）またはミス率は、誤って分類された陽性サンプルの割合です。したがって、これは感度測定を補完するものであり、式(5)で定義されています。FPRとFNRの両方ともデータ分布の変化には敏感ではなく、したがって、両方のメトリックは不均衡なデータで使用することができます[9]。

$$FPR = 1 - TNR = \frac{FP}{FP + TN} = \frac{FP}{N}$$

$$FNR = 1 - TPR = \frac{FN}{FN + TP} = \frac{FN}{P}$$

2.5. Predictive values

陽性予測値(PPV)またはPrecisionは、式（6）[20]で示されるように、陽性と予測されたサンプルの総数に対して正しく分類された陽性サンプルの割合を表している。一方、陰性予測値(NPV)、逆精度、または真陰性精度(TNA)は、式（7）[16]で示された陰性と予測された検体の総数に対する正しく分類された陰性検体の割合を

測定します。これら2つの測定値は、不均衡なデータに対して敏感である[21,9]。偽発見率 (FDR) は PPV、偽省略率 (FOR) は NPV を補完する尺度である (式(6)、(7)参照)。

$$PPV = Precision = \frac{TP}{FP + TP} = 1 - FDR$$

$$NPV = \frac{TN}{FN + TN} = 1 - FOR$$

また、精度は、以下のようにprecisionとinverse precisionで定義することができる[16]。

$$ Acc = \frac{TP+FP}{P+N} \times PPV + \frac{TN+FN}{P+N} \times NPV \ = \frac{TP+FP}{P+N} \times \frac{TP}{TP+FN} + \frac{TN+FN}{P+N} \times \frac{TN}{TN+FN} \ = \frac{TP+TN}{TP+TN+FP+FN} $$

2.6. Likelihood ratio

尤度比は、感度と特異度の両方を兼ね備えたもので、診断検査で用いられる。尤度比は、結果が確率に与える影響を測定する。陽性尤度(LR+)は、診断結果が陽性の場合に病気の確率がどの程度上昇するかを示すもので、式(9) [20]のように計算される。同様に、陰性尤度(LR-)は、診断結果が陰性の場合に病気の確率がどの程度低下するかを示すもので、式(9)のように計算される。どちらの尺度も感度と特異度の尺度に依存するので、バランスのとれたデータと不均衡なデータに適している[6]。

$$LR+ = \frac{TPR}{1 - TNR} = \frac{TPR}{FPR}$$

$$LR- = \frac{1 - TPR}{TNR}$$

LR+とLR-の両方を1つの指標にまとめ、試験の性能を要約したもので、この指標を診断オッズ比(DOR)と呼ぶ。DORは、式(10)のように正の尤度比と負の尤度比の比を表す。この指標は、テストの識別能力を推定するために、また、2つの診断テスト間の比較のために利用されます。式(10)から，(1)TPとTNが高く，(2)FPとFNが低い場合にDORの値が増加することがわかる[18].

$$DOR = \frac{LR+}{LR-} = \frac{TPR}{LR-} = \frac{TPR}{1 - TNR} \times \frac{TNR}{1 - TPR} = \frac{TP \times TN}{FP \times FN}$$

2.7. Youden's inder

ユーデンの指標（YI）またはブックメーカーの情報度（BM）メトリックは、よく知られている診断テストの一つです。それはテストの識別力を評価します。Youden's indexの式は、DORメトリックと同様に感度と特異度を組み合わせたもので、

$$YI = TPR + TNR - 1$$

[20]のように定義されています。YI メトリックは、テストが悪いときにゼロから完全な診断テストを表す 1 つまでの範囲である。また，不均衡なデータにも適している。この検査の主な欠点の1つは、検査の感度と特異度の差に関して変化しないことである。例えば，2つの検査が与えられた場合，1回目と2回目の検査の感度値がそれぞれ0.7と0.9，1回目と2回目の検査の特異度値がそれぞれ0.8と0.6の場合，両方の検査のYI値は0.5となる。

2.8. Another metrics

これまでのメトリクスから計算できるメトリクスは、さまざまなものがあります。各指標の詳細は以下の通りです。

マシューズ相関係数(MCC)：1975 年にブライアン・W・マシューズによって導入された指標である[14]。係数が+1の場合は完全な予測を示し、-1の場合は予測と真の値の不一致を表し、ゼロの場合はランダムな予測よりも優れていないことを意味します[16,3]。この指標は不均衡なデータに敏感である。

$$ MCC = \frac{TP \times TN -FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \ = \frac{\frac{TP}{N}-TPR \times PPV}{\sqrt{PPV \times TPR(1-TPR)(1-PPV)}} $$

判別力（DP）：この尺度は感度と特異度に依存し、

$$DP = \frac{\sqrt{3}}{\pi}\left(\log\left(\frac{TPR}{1-TNR}\right)+\log\left(\frac{TNR}{1-TPR}\right)\right)$$

として定義されます。[20]. この指標は、分類モデルがどれだけ陽性サンプルと陰性サンプルを区別できるかを評価する。このメトリックは感度メトリックと特異度メトリックに依存するので、不均衡なデータでも使用できる。

F値：F1スコアとも呼ばれ、(12)式[20]の真陽性精度と真陽性率の調和平均を表します。F値は0から1まであり、F値が高いほど分類性能が高いことを示します。この指標には、$F_\beta - measure$ と呼ばれる別のバリアントがあります。これは、式(13)のように、精度とリコールの間の加重調和平均を表します。

$$F-measure = \frac{2PPV \times TPR}{PPV+TPR} = \frac{2TP}{2TP+FP+FN}$$

$$F_\beta - measure = (1+\beta^2)\frac{PPV \times TPR}{\beta^2 PPV+TPR} = \frac{(1+\beta^2)TP}{(1+\beta^2)TP+\beta^2 FN+FN}$$

この尺度は、データ分布の変化に敏感です。負のクラス サンプルがα倍に増加したと仮定します。すると、F値は次のように計算され、

$$F-measure = \frac{2TP}{2TP+\alpha FP+\alpha FN}$$

となります。 [13]では調整F値(AGF)が導入されました。Fメジャーは，混同行列の4つの要素のうち3つの要素のみを使用していたため，異なるTNR値を持つ2つの分類器が同じFスコアを持つことがありました．そのため，混同行列のすべての要素を使用し，少数派クラスに正しく分類されたサンプルにより多くの重みを与えるために，AGFメトリックが導入された．このメトリックは以下のように定義されます。

$$AGF = \sqrt{F_2 InvF_{0.5}}$$

ここで、F_2は、$\beta = 2$の場合の$F-measure$であり、$InvF_{0.5}$は、各サンプルのクラスラベルが入れ替わる（すなわち、正のサンプルは負のサンプルになり、逆の場合もある）新しい混同行列を構築することによって計算されます。

マークドネス(MK)：PPVとNPVに基づいて定義され、

$$MK = PPV+NPV-1$$

[16]のようになる。このメトリックは、データの変化に敏感であるため、不均衡なデータには適していない。これは、マークドネスメトリックがPPVとNPVに依存しており、PPVとNPVの両方がデータ分布の変化に敏感であるためである。

バランス分類率またはバランス精度(BCR)：感度と特異度を組み合わせたもので、

$$BCR = \frac{1}{2}(TPR + TNR) = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$$

のように計算されます。また、バランスエラー率(BER)またはハーフトータルエラー率(HTER)は、$1 - BCR$を表しています。BCRとBERの両方のメトリックは、不均衡なデータセットで使用することができます。

幾何平均(GM)。**すべての分類器の主な目標は，特異性を犠牲にすることなく感度を向上させること**である．しかし、感度と特異度の目的はしばしば相反するものであり、特にデータセットが不均衡な場合には、うまく機能しないことがあります。そこで、Geometric Mean(GM)メトリックは、式(15) [3]に従って感度と特異度の両方の尺度を集約します。各クラスについて可能な限り多くの情報を得るために、調整幾何平均(AGM)が提案されている[11]。AGMメトリックは、式(16)に従って定義される。

$$GM = \sqrt{TPR \times TNR} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}}$$

AGM = \left { \begin{array}{ccc} \frac{GM+TNR(FP+TN)}{1+FP+TN} & if & TPR > 0 \ 0 & if & TPR = 0 \end{array} \right.

GMメトリックは不均衡なデータセットでも使用できる。Lopezらは、AGMメトリックが不均衡データに適していることを報告している[12]。しかし，負のクラスの分布の変化がAGMメトリックに小さく影響を与えるため，したがって，不均衡なデータには適していない．これは、負のクラスのサンプルをα倍に増やすと仮定するだけで証明できます。このようにして、AGM メトリックは以下のように計算されます,

$$AGM = \frac{GM + TNR(\alpha FP + \alpha TN)}{1 + \alpha FP + \alpha TN}$$

結果として、AGM メトリックはクラス分布の変化の影響をわずかに受けます。

最適化精度(OP)。このメトリックは次のように定義されます。

$$OP = Acc - \frac{|TPR - TNR|}{TPR + TNR}$$

ここで2番目の項$\frac{|TPR-TNR|}{TPR+TNR}$は両クラスの精度がどれだけバランスが取れているかを計算しており、このメトリックはグローバル精度とその項の差を表している[9]。OP値が高いほど精度が高く、クラス精度のバランスが取れていることを示します。OP値は精度に依存するため、不均衡なデータには不向きである。

ジャカード。このメトリックは谷本類似度係数とも呼ばれる。Jaccardメトリックは、以下のように、

$$Jaccard = \frac{TP}{TP + FP + FN}$$

のように、負の標本の正しい分類を明示的に無視します。Jaccard metricはデータ分布の変化に敏感である。

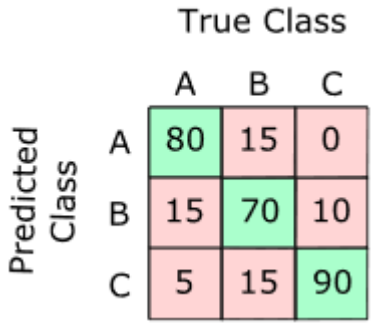
2.9. Illustrative example

このセクションでは、2つの例を紹介します。これらの例は、2 つのクラスまたは複数のクラスを使用して、分類メトリックを計算する方法を説明します。

2.9.1 Binary classification example.

この例では、2つのクラス（AとB）、すなわち2値分類があり、各クラスのサンプル数が100個であると仮定します。Aクラスは正のクラスを表し、Bクラスは負のクラスを表します。AクラスとBクラスの正しく分類されたサンプル数はそれぞれ70個と80個です。したがって、TP；TN；FP；FNの値は、それぞれ70、80、20、30である。異なる分類メトリックの値は以下の通りである。

2.9.2. Multi-classification example.



この例では、A,B,Cの3つのクラスがあり、分類テストの結果を図4に示します。図から、TP_A、TP_B、TB_Cの値はそれぞれ80、70、90となり、[図4]の対角線を表している。各クラス（真のクラス）の偽陰性の値は、前述したように、そのクラスの列にある全ての誤差を加算して算出します。例えば、$FN_A = E_{AB} + E_{AC} = 15 + 5 = 20$, 同様に $FN_B = E_{BA} + E_{BC} = 15 + 15 = 30$ と $FN_C = E_{CA} + E_{CB} = 0 + 10 = 10$ です。各クラス(予測されたクラス)の偽陽性の値は、前述のように、そのクラスの行にあるすべての誤差を加算して計算されます。例えば、$FP_A = E_{BA} + E_{CA} = 15 + 0 = 15$, 同様に $FP_B = E_{AB} + E_{CB} = 15 + 10 = 25$, と $FP_C = E_{AC} + E_{BC} = 5 + 15 = 20$ です。クラスAの真の負の値(TN_A)は、クラスAの行および列を除くすべての列および行を加算することによって計算することができる；これは、2×2 混同行列のTNと同様である。したがって、TN_Aの値は次の+ように計算される。$TN_A = 70 + 90 + 10 + 15 = 185$, と同様に $TN_B = 80 + 0 + 5 + 90 = 175$ と $TN_C = 80 + 70 + 15 + 15 = 180$ です。 TP, TN, FP, FNを使用すると、すべての分類尺度を計算することができます。例えば、Accuracyは

$$Acc = \frac{80 + 70 + 90}{100 + 100 + 100}$$

です。クラスごとの感度と特異度、例えば、Aの感度は

$$TPR_A = \frac{TP_A}{TP_A + FN_A} = \frac{80}{80 + 15 + 5} = 0.8$$

と同様に、BクラスとCクラスの感度はそれぞれ

$$TPR_B = \frac{TP_B}{TP_B + FN_B} = \frac{70}{70 + 15 + 15} = 0.7$$

と

$$TPR_C = \frac{TP_C}{TP_C + FN_C} = \frac{90}{90 + 0 + 10} = 0.9$$

です。 また、 A、 B、 Cの特異度の値はそれぞれ

$$TNR_A = \frac{TN_A}{FP_A + TN_A} = \frac{185}{15 + 185} \approx 0.93$$

$$TNR_B = \frac{TN_B}{FP_B + TN_B} = \frac{175}{25 + 175} \approx 0.875$$

$$TNR_C = \frac{TN_C}{FP_C + TN_C} = \frac{180}{20 + 180} \approx 0.9$$

3. Receiver operating characteristics (ROC)

受信機動作特性（ROC）曲線は、TPR を y 軸、FPR を x 軸とした 2 次元グラフです。ROC曲線は、診断システム、医療意思決定システム、機械学習システムなど多くのシステムの評価に使用されています[26]。これは、利点、すなわち真の陽性とコスト、すなわち偽陽性の間のバランスをとるために使用される。決定木のような離散的な出力を持つ分類器は、クラス決定、すなわち各テストサンプルに対する決定のみを生成するように設計されており、したがって、ROC空間への1点に対応する1つの混同行列のみを生成する。しかし，分類器から完全なROC曲線を生成するために，1点だけではなく，クラスの割合を使用したり，スコアリングと投票の組み合わせを使用したりするなど，多くの方法が紹介されています[26]．一方，Naive Bayes 分類器のような連続出力の分類器では，出力は数値，すなわちスコアで表され，標本が特定のクラスにどの程度属しているかを表します．ROC曲線は，信頼度スコアのしきい値を変化させることで生成されます；したがって，各しきい値はROC曲線の1点のみを生成します[8]．

7. Experimental results

本節では、異なる評価方法を用いて分類性能を評価する実験を行いました。この実験では、カリフォルニア大学アービン校(UCI)の機械学習リポジトリ[1]から入手した、標準的な分類データセットの一つであるアイリスデータセットを用いました。このデータセットは3つのクラスを持ち、各クラスは50個のサンプルを持ち、各サンプルは4つの特徴量で表現されています。特徴量を2つに減らすために(1)主成分分析(PCA) [23]を、分類には(2)サポートベクターマシン(SVM)を用いた。

本実験では，学習モデルを評価するために異なる評価方法を用いた．図13はROC曲線とPrecision-Recall曲線を示している．曲線は3つあり、各クラスごとに1つの曲線があり、このように1つ目のクラスが他の2つのクラスよりも良い結果を得ています。図14は、各クラスの混同行列を示す。これらの混同行列から、先に述べたように、さまざまな指標を算出することができます（図3参照）。例えば、最初のクラスの結果は、Acc、TPR、TNR、PPV、NPVは、それぞれ99.33、100、98.0、99.01、100であった。同様に、他の 2 つのクラスの結果も算出することができる。